Développement d'outils de géo-calcul haute performance pour l'identification de régions du génome potentiellement soumises à la sélection naturelle: analyse spatiale de la diversité de panels de polymorphismes nucléotidiques à haute densité (800k) chez *Bos taurus* et *B. indicus* en Ouganda

THÈSE N^{O} 6014 (2014)

PRÉSENTÉE LE 28 FÉVRIER 2014

À LA FACULTÉ DE L'ENVIRONNEMENT NATUREL, ARCHITECTURAL ET CONSTRUIT
LABORATOIRE DE SYSTÈMES D'INFORMATION GÉOGRAPHIQUE
PROGRAMME DOCTORAL EN ENVIRONNEMENT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Sylvie STUCKI

acceptée sur proposition du jury:

Prof. A. Buttler, président du jury
Prof. F. Golay, Dr S. Joost, directeurs de thèse
Prof. M. W. Bruford, rapporteur
Prof. S. Morgenthaler, rapporteur
Dr P. Taberlet, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

Ainsi donc, si vous le voulez bien…
Par avance je vous en remercie.
Ne vous affolez pas et ne craignez rien.
Ne buvez que de l'eau, rien d'autre,
sinon vous vous ramolliriez et vous
auriez du mal à tenir le coup.
C'est l'heure !
— Woland

(Mikhaïl Boulgakov,
Le Maître et Marguerite)

A mes parents
A Patrick

Remerciements

Je tiens à exprimer ma plus chaleureuse gratitude à mes directeurs de thèse, le Prof. François Golay et le Dr. Stéphane Joost, pour la confiance qu'ils m'ont accordée et leur soutien indéfectible tout au long de mon parcours. Stéphane, merci tout particulièrement pour ta disponibilité, pour tes explications et tes précieux conseils ainsi que pour ton immense travail de relecture du manuscrit.

Je souhaite remercier les institutions qui m'ont soutenue : la Commission Européenne au travers du projet FP7 NextGen, la *European Science Foundation* et La Poste Suisse.

De nombreuses personnes ont joué un rôle dans la réussite de ce travail. Merci aux membres du LASIG, présents et anciens, pour notre excellente atmosphère de labo, en particulier Kevin, Tim, Devis, Ema, Véronique, Matthew, Marc, Ivo, Jens, Michaël, Ahmed, Nicolas, Jessie, Magda et Julien. J'aimerais remercier Prof. Michel Bruford, Prof. Sergio Rey, Dr. Pierre Taberlet, Prof. Stephan Morgenthaler, Eric Frichot, Prof. Graham Coop, Dr. François Pompanon, Dr. Licia Colli pour leurs conseils et leur explications. Merci également à Jean-Daniel Bonjour, David Meylan et Samuel Bancal pour leur aide dans mes tâches informatiques.

J'aimerais remercier tous les membres du projet NextGen qui ont rendu cette expérience à la fois plaisante et enrichissante. *Muchas gracias* à Pablo pour son soutien à la fois scientifique et personnel.

Je tiens à remercier ma famille et mes amis pour leurs constants encouragements et en particulier mes parents, André et Elisabeth, qui m'ont permis d'arriver jusqu'ici. J'aimerais remercier Nadège qui est toujours aussi proche malgré la distance qui nous sépare, et également Diane, Marc et Blaise, mes compagnons de route à l'EPFL, qui savent toujours me faire rire.

J'ai connu des moments plus difficiles que d'autres, mais la vie est toujours plus agréable à deux. Je remercie plus que je ne puis le dire Patrick qui me fait l'immense joie d'être tous les jours à mes côtés.

Lausanne, le 24 octobre 2013 Sylvie Stucki

Résumé

Le recours aux techniques de séquençage du génome entier permet désormais d'étudier la variabilité génétique d'une population avec une précision inédite. Toutefois la taille des jeux de données disponibles présente un défi pour leur analyse. En particulier, les études concernant les bases génétiques de l'adaptation locale doivent s'adapter à ce nouveau contexte et c'est le cas notamment des approches corrélatives en génomique environnementale.
Dans le cadre du projet européen NextGen (nextgen.epfl.ch), dédié à la mise au point de stratégies de conservation des ressources génétiques de races locales d'animaux domestiques, cette thèse vise principalement à développer une approche corrélative rapide pour détecter des régions du génome potentiellement soumises à la sélection naturelle dans de grands jeux de données moléculaires. Ce type d'approche modélise la variabilité génétique des organismes étudiés à partir des caractéristiques environnementales de leurs habitats, ce qui permet de formuler des hypothèses sur les processus écologiques influençant l'adaptation locale. Cependant les approches corrélatives sont enclines à faire de fausses découvertes dans les cas où les organismes étudiés présentent une structure de population due à leur histoire démographique.
Cette structure est le résultat de la transmission des gènes entre individus en fonction de la localisation et de leurs déplacements dans l'espace géographique (dispersion). Les statistiques spatiales mises en œuvre dans cette thèse offrent justement la possibilité de mesurer l'autocorrélation spatiale des fréquences alléliques en présence, et peuvent être utilisées pour limiter la production de faux positifs évoqués plus haut.
Dans cette recherche, les résultats produits par le logiciel développé appelé Samβada [1] sont comparés avec ceux générés par deux autres méthodes corrélatives qui intègrent la démographie (`BayEnv` et `LFMM`) ainsi qu'avec les résultats produits par une approche de génétique des populations (`Arlequin` — approche `F-Dist`). Les quatre méthodes s'accordent pour désigner un groupe de trois loci situés sur le chromosome 5 près du gène « BT.42818 » chez les bovins étudiés. Ce gène est impliqué dans la synthèse d'une protéine dont l'homologue chez l'être humain joue un rôle dans la détection et l'élimination des bactéries intracellulaires. Or les bovins porteurs de ces mutations se trouvent dans une région touchée par la trypanosomiase (maladie du sommeil). Ces indications suggèrent que les loci détectés sont impliqués dans un processus de résistance aux parasites. Plus généralement ces résultats intéressants sont le

1. Le logiciel est disponible en ligne : lasig.epfl.ch/sambada

produit de nombreux calculs appliqués à un gros volume de données moléculaires (600'000 SNPs). Ils démontrent que le logiciel proposé implémente une méthode très rapide, capable de traiter prochainement des données représentant le génome entier.

Mots-clés : génomique environnementale, adaptation locale, calcul haute performance, approches corrélatives, autocorrélation spatiale, systèmes d'information géographique, séquençage du génome entier, génétique des populations, résistance aux parasites.

Abstract

Whole-genome sequencing techniques open the possibility to study the genetic variability in a population with an unprecedented precision. However the size of the available datasets is a challenge. Studies of local adaptation have to adapt to this new context and in particular the correlative approaches in landscape genomics have to be modified.
In the framework of the NextGen European project (nextgen.epfl.ch) dedicated to designing conservation strategies for preserving the genetic diversity of local breeds of farm animals, this thesis focuses on developing a computationally-fast solution for detecting genomic regions possibly subject to natural selection. We model the genetic variability of the organisms of interest taking into account the environmental characteristics of their habitats. This allows us to formulate hypotheses regarding the ecological processes influencing local adaptation. Correlative approaches tend to make false discoveries when the organisms under study present some additional structure due to their demographic history.
This results from gene transmission between individuals based on their location and movements in the geographical space (dispersion). Spatial statistics is used in this thesis in order to measure spatial autocorrelation of current allelic frequencies, and can be used to limit the occurence of false positives as mentioned above.
In our study, results provided by this newly-developped software called Samβada [2] are compared with those generated by two other programs that integrate demography (`BayEnv` and `LFMM`) and the use of population genetics (`Arlequin` — `F-Dist` method). The four methods agree on detecting a group of three loci located on chromosome 5 close to the gene "BT.42818" in the cattle under study. This gene has a function in the synthesis of a protein whose human counterpart has a role in detecting and eliminating intracellular bacteria. The bovines carrying this mutation live in a region affected by trypanosomiasis (sleeping sickness). These pieces of information suggest that these loci are involved in parasite resistance. These results constitute the outcome of numerous computations on a large molecular dataset (600'000 SNPs). Comparison of the performances shows that the new software implements a very fast method, able to process whole-genome data efficiently.
Keywords: landscape genomics, local adaptation, high performance computing (HPC), correlative approaches, spatial autocorrelation, geographic information systems, whole-genome sequencing, population genetics, resistance to parasites

2. The software is available at lasig.epfl.ch/sambada

Table des matières

Table des figures

Liste des tableaux

1 Introduction

En écologie moléculaire, la récente disponibilité de données génomiques haute densité a ouvert de nouvelles perspectives dans tous les domaines de la discipline, notamment en GÉNOMIQUE ENVIRONNEMENTALE [1] (*Landscape genomics,* Luikart et al., 2003 ; Joost et al., 2007 ; Manel et al., 2010). Les études menées il y a une dizaine d'années portaient en moyenne sur une trentaine de marqueurs génétiques (Manel et al., 2003) et ce nombre a progressivement augmenté jusqu'à quelques centaines suivant les techniques utilisées (Poncet et al., 2010). L'avènement actuel du séquençage à haut débit (Kijas et al., 2012), qui permet de déchiffrer le GÉNOME entier (The 1000 Genomes Project Consortium, 2012), fournit une image de la diversité génétique des organismes étudiés avec une précision inédite. Cette évolution a notamment permis au projet européen NextGen [2] de séquencer 450 animaux de ferme pour lesquels 16 millions de variations génétiques ont été identifiées.

Les méthodes d'analyse utilisées jusqu'ici doivent être adaptées afin de pouvoir traiter de grands jeux de données moléculaires. La génomique environnementale, où la fréquence d'apparition des variantes génétiques est modélisée en relation avec l'environnement, est également concernée, et c'est dans ce domaine que s'inscrivent les travaux de recherche qui constituent cette thèse. Il s'agit ici de mettre en œuvre des méthodes corrélatives spécialement conçues pour traiter de grands jeux de données moléculaires afin de détecter les régions adaptatives du génome chez les animaux domestiques, de simplifier et d'automatiser le traitement et la visualisation des résultats et d'utiliser l'analyse spatiale pour enrichir leur interprétation.

La science de l'information géographique joue un rôle central en génomique environnementale puisqu'elle permet de décrire l'habitat des organismes étudiés. La connaissance de la localisation des individus révèle aussi la distribution spatiale de la diversité génétique. L'analyse spatiale permet notamment de mesurer la similitude entre individus en fonction de la distance les séparant et d'en déduire diverses informations liées à leur processus de reproduction et à leur interaction avec l'environnement.

1. Les termes en petites capitales sont définis dans le glossaire.
2. cf nextgen.epfl.ch et p. 5.

1.1 Méthodes de détection de régions du génome soumises à la sélection naturelle

La génomique environnementale met principalement en oeuvre des techniques de statistiques corrélatives qui permettent - lorsque les coordonnées géographiques des individus échantillonnés sont connues - de mesurer l'intensité et la significativité de la relation entre la fréquence de certains marqueurs moléculaires et les caractéristiques de leur environnement local. Cela permet d'identifier des régions du génome (des LOCI) potentiellement soumises à la SÉLECTION NATURELLE, et par conséquent de fournir des informations sur les mécanismes moléculaires qui contrôlent les processus fondamentaux liés à l'évolution des espèces. Une étude récente de deux espèces de coraux de la Grande Barrière le long d'un gradient environnemental a par exemple permis de détecter des MUTATIONS génétiques liées à la résistance au stress induit par la variation de la température et celle de la turbidité de l'eau (Lundgren et al., 2013).

Cependant, c'est la GÉNÉTIQUE DES POPULATIONS qui a fourni la première des approches capables d'identifier les signatures de sélection. Ces méthodes sont généralement basées sur la détection de loci singuliers (*outliers*). Elles supposent que les loci soumis à la sélection naturelle présentent un comportement distinct de celui des mutations neutres (par rapport à la sélection naturelle), qui forment l'essentiel de la variabilité génétique (Schoville et al., 2012). Ces approches requièrent que la structure de population sous-jacente soit correctement prise en compte pour éviter de faire de fausses découvertes. Les méthodes de simulations coalescentes modélisent la transmission héréditaire de l'information génétique sur de nombreuses générations. Elles permettent de déterminer le comportement des loci neutres au prix d'une augmentation du temps de calcul.

En comparaison, un avantage important de la génomique environnementale est qu'elle permet d'identifier les loci soumis à la sélection naturelle tout en fournissant des informations sur les processus écologiques à l'œuvre. Elle part du principe que les individus doivent être étudiés en relation avec leur habitat car leur environnement local exerce une pression de sélection sur eux (cf fig. 1.1). Au cours des générations, les individus répondent à cette pression soit grâce à une plasticité PHÉNOTYPIQUE, leur morphologie étant capable de varier en fonction des conditions écologiques, soit en s'adaptant aux conditions locales. Dans le deuxième cas, leur patrimoine génétique est modifié, de nouvelles mutations apparaissent ou la fréquence des mutations existantes change. La fraction de la variabilité génétique qui est impliquée dans l'adaptation reflète donc les conditions environnementales du milieu (fig. 1.1, flèche 5). La génomique environnementale modélise la probabilité qu'un individu porte un marqueur génétique en fonction des caractéristiques de son habitat. Elle suppose que la sélection naturelle a eu le temps de créer un gradient de fréquences alléliques parmi les représentants de son espèce et que la relation génome-environnement est constante dans l'espace. Les méthodes corrélatives qui n'utilisent pas de modèle biologique ont tendance à détecter des faux positifs en cas de structure spatiale de population. A contrario, la détection de loci singuliers en génétique des populations se base sur des modèles démographiques,

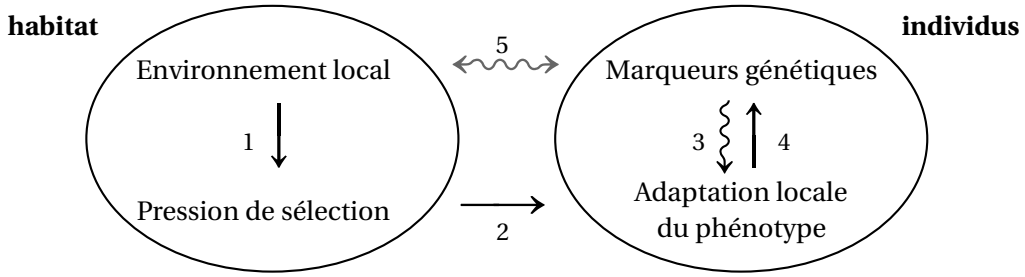


Figure 1.1 – Principe de la génomique environnementale. L'environnement exerce une pression de sélection sur les organismes (1). En parallèle, les variations du patrimoine génétique (dont l'apparition de nouvelles mutations) créent de la diversité phénotypique entre individus (3). Sous l'effet de la sélection naturelle, certains phénotypes sont favorisés au cours des générations, menant à l'adaptation locale (2). Le processus d'adaptation influence en retour le patrimoine génétique : les GÉNOTYPES produisant les phénotypes les plus adaptés se généralisent, d'autres génotypes disparaissent (4). Par conséquent, l'environnement finit par façonner le génome et les populations s'adaptent différemment suivant leur habitat. La génomique environnementale étudie la probabilité qu'un organisme porte un marqueur génétique en fonction son environnement local (5).

mais ne permet pas de formuler des hypothèses écologiques. Les fausses découvertes y sont cependant moins nombreuses, à condition de considérer le bon modèle sous-jacent (Li et al., 2012).

Il y a quelques années, les méthodes corrélatives devaient mettre en relation des jeux de données de l'ordre d'une centaine de régions du génome avec quelques dizaines de variables environnementales ; elles sont désormais couramment appelées à traiter des centaines de milliers de loci. De plus, le séquençage intégral du génome permet de déchiffrer les quelques millions de loci potentiellement variables. Il est donc nécessaire que les outils de la génomique environnementale soient adaptés à ce nouveau contexte, notamment en intégrant des techniques de calcul haute performance.

1.2 Génomique environnementale et séquençage du génome entier

Des méthodes corrélatives performantes capables d'effectuer des calculs de cette ampleur dans des temps raisonnables n'existaient pas jusque-là. Le présent travail répond à ce besoin de performance et s'inscrit dans le cadre du projet de recherche NextGen qui utilise le GÉNOTYPAGE À HAUTE DENSITÉ et le SÉQUENÇAGE DU GÉNOME ENTIER pour étudier l'adaptation locale et la résistance aux parasites chez trois espèces d'animaux domestiques en Afrique. NextGen cherche à développer des méthodes d'élevage durables et à conseiller les éleveurs locaux en vue de préserver les ressources génétiques animales (*Farm Animal Genetic Resources*, *FAnGR*).

La tâche en question présente trois types de contraintes :

Charge de calculs La méthode proposée doit permettre de calculer rapidement de nombreux modèles d'association entre le génome (∼16 millions de marqueurs) et l'environnement.

Au vu de la taille des données attendues, la version de bureau doit être accompagnée d'une version minimale et rapide dédiée au calcul haute performance sur des grappes d'ordinateurs. La méthode doit également permettre de trier et de visualiser synthétiquement et rapidement les grandes quantités de résultats.

Analyses intégrées En plus des modèles univariés qui sont généralement appliqués, l'outil doit inclure l'analyse multivariée de manière à prendre en compte de potentiels effets cumulés. La méthode doit également permettre de mesurer l'autocorrélation spatiale, qui décrit les similitudes entre points voisins, et de tenir compte des défauts connus des approches corrélatives. Ces dernières ont en effet de la difficulté à déceler les nouveaux loci adaptatifs pendant qu'ils se propagent dans la population et ont aussi tendance à détecter des faux positifs.

Accessibilité Afin d'être utile à un panel d'utilisateurs aussi large que possible, l'outil proposé doit être facile à utiliser et disponible sur différentes plate-formes. Un programme libre à code ouvert (*open source*) est par définition accessible à tous et permet également à la communauté des utilisateurs de l'améliorer.

Ces développements sont appliqués ici à l'analyse spatiale de la diversité de panels de POLYMORPHISMES NUCLÉOTIDIQUES à moyenne (50'000 loci, « 50k ») et à haute densité (800'000 loci, « 800k ») chez *Bos taurus* et *B. indicus* en Ouganda. Les animaux domestiques que nous connaissons aujourd'hui sont issus d'un long processus initié par les premiers éleveurs.

1.3 Agriculture et sélection humaine

Depuis leur domestication au Néolithique (entre 8'000 et 10'000 ans av. J.-C.), les animaux domestiques et les plantes cultivées ont été soumis à l'effet conjoint des sélections naturelle et humaine et ont ainsi lentement évolué en de nombreuses variétés et races localement adaptées (Gepts et Papa, 2003). Ces spécialisations incluent l'accoutumance aux climats extrêmes, la tolérance au manque de ressources et la résistance aux maladies et aux parasites (Mirkena et al., 2010). Cependant, depuis le milieu du XX^{e} s., la théorie génétique de l'hérédité et les progrès technologiques ont permis le développement d'une sélection intensive des animaux domestiques. Ces nouvelles méthodes d'élevage ont accru la productivité, mais également les besoins de ces animaux en nourriture et médicaments. Grâce à l'insémination artificielle, les individus ayant la plus haute valeur génétique (*breeding value*) ont de nombreux descendants. Bien que ces races cosmopolites soient très répandues dans les élevages industriels, les tailles effectives des populations sont petites sous l'effet de la consanguinité (Taberlet et al., 2008). Les ressources génétiques nécessaires à l'adaptation à de nouvelles conditions ne sont que rarement prises en compte dans les croisements car elles n'augmentent pas la valeur marchande de l'animal. Dans le domaine de la production laitière, les nouvelles techniques de génotypage à haute densité permettent d'estimer la valeur génétique d'un mâle sans attendre de tester sa descendance (Goddard et al., 2010), ce qui accélère le processus de sélection et l'uniformisation des élevages.

Les races traditionnelles sont aujourd'hui menacées par la concurrence de ces races cosmopolites hautement productives mais très homogènes. Certaines races locales ont déjà disparu (Hoffmann, 2011). Les croisements entre races locales et cosmopolites, qui visent à augmenter la productivité des premières, abaissent leur variabilité génétique et mettent parfois en péril leur descendance, notamment dans les pays en voie de développement où les animaux ne peuvent recevoir de soins médicaux (Berthouly-Salazar et al., 2012). Ces croisements sont causés par l'accroissement des besoins humains, mais ne tiennent souvent pas compte de la rudesse des conditions locales, accentuée par les changements climatiques, ni surtout du fait que les races locales possèdent justement les ressources génétiques leur permettant de s'adapter (Hanotte et al., 2010). Le développement d'une agriculture durable nécessite de répertorier et d'étudier la diversité génétique des animaux domestiques et des plantes cultivées afin de conserver leur variabilité. Certains programmes d'élevage permettent déjà de prendre en compte la robustesse des animaux (Bett et al., 2012). Face à l'augmentation des besoins et à la pression exercée sur les races traditionnelles, l'objectif de NextGen est d'étudier les bases génomiques de l'adaptation locale pour développer des méthodes d'élevage qui préservent les ressources génétiques des animaux domestiques.

1.4 Régions et races d'étude

L'Ouganda, pays d'Afrique de l'Est, recèle deux principales espèces de bovins (cf fig. 4.11 p. 50). Les vaches ankoles (*B. taurus*), venues d'Europe aux alentours de 3000 avant J.-C., peuplent le sud-ouest du pays, tandis que les zébus (*B. indicus*) ont été introduits depuis l'Inde vers 700 après J.-C. et sont élevés dans le nord-est du pays (Ajmone Marsan et al., 2010). La distribution des deux espèces est liée à la présence du trypanosome, parasite transmis par la mouche tsé-tsé, qui provoque la maladie du sommeil. Les ankoles semblent plus résistantes à ce parasite que les zébus, ce qui limite la dispersion de ces derniers (Groeneveld et al., 2010).

L'introgression d'animaux cosmopolites est rapportée dans plusieurs régions (fig. 1.2). La vache holstein est une race laitière très productive mais très homogène et peu résistante aux stress environnemental (Taberlet et al., 2011). Les croisements visent à augmenter la productivité des vaches locales, au prix d'une perte de robustesse. A terme, les hybrides risquent de perdre leur faculté de résister au trypanosome. L'étude de la résistance aux parasites fait partie du projet et les données récoltées permettent notamment de tester les développements des nouveaux algorithmes corrélatifs. La détection de signatures de sélection dans le génome des zébus et des ankoles cherche principalement à dévoiler les régions du génome impliquées dans la tolérance à la charge parasitaire.

La partie du projet dédiée à l'étude de l'adaptation locale est principalement menée au Maroc. Ce pays possède une grande diversité de conditions environnementales, entre montagnes, plaines et désert. Les chèvres et les moutons y vivent dans des habitats très contrastés et l'analyse de leur variabilité génétique vise à détecter les régions génomiques sous-tendant l'adaptation locale. NextGen permet pour la première fois un séquençage intégral et une

(a) Vaches ankoles dans la région de Mbarara

(b) Vache holstein en France

(c) Vache hybride dans la région de Mbarara

Figure 1.2 – Exemple d'introgression de bovins en Ouganda. Les holsteins sont des vaches laitières très courantes en Europe et sont souvent exportées pour des croisements visant à augmenter la productivité de races locales. Les hybrides ont rarement la robustesse de leurs ancêtres locaux[3].

analyse en relation avec l'environnement pour un grand nombre de moutons et de chèvres. La taille des jeux de données attendus est à l'origine de mes travaux de recherche et a motivé mes développements. Toutefois, la lecture de l'information génétique est actuellement beaucoup plus rapide avec le génotypage qu'avec le séquençage. Les données produites par le génotypage des vaches ougandaises ont donc précédé celles des ruminants marocains et c'est pourquoi les analyses présentées au chapitre 7 concernent les zébus et les ankoles.

Finalement c'est parce que l'Iran est un des berceaux de la domestication des mouflons (*Ovis orientalis*) et des bézoars (*Capra aegagrus*), que NextGen y étudie les ancêtres sauvages du mouton et de la chèvre dans le but d'analyser les ressources génétiques disponibles. En effet les animaux vivant près des centres de domestication présentent une plus grande variabilité génétique que ceux qui en sont éloignés (Groeneveld et al., 2010). Ces animaux pourraient participer à des programmes d'élevage pour accroître la taille effective de population dans d'autres races. La diversification du patrimoine génétique des chèvres et des moutons domestiques leur donnerait la variabilité nécessaire pour s'adapter à de nouvelles conditions environnementales. Cette approche nécessite des animaux vivants, elle ne peut s'appliquer à

3. Illustrations : (a) et (c) Gentile, 2011 ; (b) Tractorboy60, 2004.

la vache car son ancêtre sauvage, l'auroch, a disparu au XVII^{e} s.

Les races d'animaux d'élevage autochtones sont un modèle de choix pour l'étude de la sélection naturelle et les nouvelles techniques de séquençage fournissent une vision détaillée de leur ressources génétiques. L'enjeu est de développer des méthodes d'analyse permettant d'appréhender ces jeux de données.

2 Détection de la sélection naturelle dans les grands jeux de données moléculaires

Le domaine de la détection de la sélection naturelle dans le patrimoine génétique est en plein essor notamment grâce aux techniques de génotypage et de séquençage à haut débit. Dans cette partie, je présente les enjeux actuels de la génomique environnementale liés aux gros volumes de données, ainsi que les objectifs de ma recherche visant à améliorer les performances des approches corrélatives.

La détection de régions génomiques soumises à la sélection connaît une expansion rapide dans le sillage des nouvelles méthodes de séquençage à haut débit (Andrew et al., 2013). En effet, les séquenceurs actuels déchiffrent plusieurs milliards de paires de bases en une journée (Liu et al., 2012). Pour une nouvelle espèce, le facteur limitant la production de connaissances nouvelles n'est plus l'extraction des données mais leur traitement (Nature editorial, 2013).

Comme mentionné plus haut, la génétique des populations propose de nombreuses méthodes pour repérer les loci soumis à la sélection parmi les loci neutres. Ces approches sont principalement basées sur la variation des fréquences alléliques, soit entre populations, soit sur le génome aux abords d'une région sous sélection (Nielsen et al., 2005). Leur principale limitation dans le traitement de grands jeux de données est le temps nécessaire pour calculer les simulations coalescentes décrivant le comportement des loci neutres. Quant à elles, les méthodes dites corrélatives utilisées en génomique environnementale prennent en compte la position des individus et la composition environnementale de leur habitat dans l'analyse des données génétiques. Elles sont plus directes et permettent de relever plus efficacement le premier défi méthodologique constitué par la taille des jeux de données.

Le second défi méthodologique pour la génomique environnementale est la capacité à intégrer les effets démographiques qui influencent la variabilité génétique. Lorsque la distance parcourue par un individu pour trouver un partenaire est petite par rapport à l'étendue géographique de son espèce, les effets des mutations et de la DÉRIVE GÉNÉTIQUE ne sont pas compensés par la dispersion. Il s'ensuit que les fréquences alléliques peuvent diverger pour former des

gradients entre les régions éloignées, c'est l'isolation par la distance. Si les habitats varient de la même façon, des corrélations fallacieuses apparaîtront entre les fréquences alléliques et les variables environnementales. La démographie peut ainsi produire des signaux similaires à la sélection et c'est pourquoi les méthodes corrélatives ont tendance à détecter des faux positifs.

Plusieurs méthodes de génomique environnementale tiennent compte de ces effets démographiques. `BayEnv` simule une distribution neutre des fréquences alléliques à partir de la répartition des individus en populations pour corriger les tests de significativité (Coop et al., 2010), tandis que `LFMM` cherche à modéliser en même temps la structure de population et l'effet de l'environnement (Frichot et al., 2013). L'hétérogénéité spatiale et l'histoire d'une population influencent la manière dont les gènes s'y diffusent ; des mutations favorisées par la sélection peuvent disparaître si les mouvements migratoires entravent leur propagation .

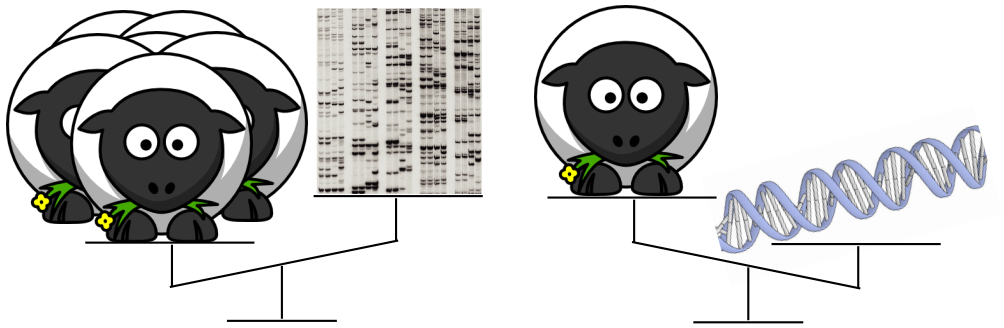
Mes recherches veulent relever en priorité le premier défi : le traitement efficace de grands jeux de données. Étant donné la complexité des modèles tenant compte de la démographie, leur inclusion dans les développements initiaux aurait considérablement ralenti l'analyse, ce qui aurait compromis la réussite de l'objectif principal. C'est pourquoi j'ai décidé de ne pas inclure ces modèles dans la méthode adoptée et de proposer à la place une approche diagnostique basée sur les mesures d'autocorrélation spatiale des loci potentiellement sous sélection.

2.1 Approches corrélatives et taille des données

Les techniques de séquençage de deuxième génération, aussi appelées *next-generation sequencing*, ont démocratisé l'accès aux données du génome entier. Le décryptage du patrimoine génétique est désormais beaucoup plus rapide et moins cher. Le projet de séquençage du génome humain, qui utilisait la méthode Sanger (première génération), dura 13 ans et impliqua 20 laboratoires pour un coût de 2,7 millards $ [1]. Cette entreprise titanesque a grandement facilité l'analyse d'autres espèces, par exemple le génome du cheval, séquencé et assemblé en une année. Le projet « 1000 Genomes » est parvenu à analyser 1'092 génomes humains avec une combinaison de méthodes en 4 ans (The 1000 Genomes Project Consortium, 2012) et NextGen a séquencé et assemblé le génome de 450 animaux domestiques en une année.

Les polymorphismes nucléotidiques (*single nucleotide polymorphisms*, SNP) sont couramment utilisés pour la détection de signatures de sélection. Un SNP est la mutation du nucléotide présent à un locus. Le séquençage et la comparaison du génome entier de plusieurs individus permettent de détecter ces modifications. On considère qu'il y a environ 1 SNP pour 1'000 paires de bases, ce qui représente entre 1 et 10 millions de SNPs pour un génome de mammifère. Les espèces les plus étudiées, comme l'homme et certains animaux d'élevage, disposent de puces de génotypage qui permettent de déchiffrer entre 3k et 800k SNPs prédéfinis à moindre coût. Un grand nombre d'individus peuvent ainsi être analysés rapidement,

1. http ://www.genome.gov/11006943



(a) Avant : beaucoup d'individus et peu de marqueurs génétiques.

(b) Après : beaucoup de marqueurs génétiques et peu d'individus.

Figure 2.1 – L'émergence du séquençage intégral du génome a provoqué un changement de paradigme pour les études génétiques [3].

mais les marqueurs ne peuvent pas être choisis. Une société privée propose même au public de génotyper 650 SNPs pour 100 $ [2].

En parallèle, la disponibilité de variables environnementales a également augmenté ces dernières années. Plusieurs bases de données climatiques sont librement accessibles en ligne (New et al., 2002 ; Hijmans et al., 2005) et les modèles numériques d'altitude à haute résolution permettent d'utiliser plusieurs échelles lors de l'analyse (Farr et al., 2007). A l'instar des données moléculaires, le nombre de variables environnementales influence la charge de calcul dans le cadre des approches corrélatives.

En plus du traitement de grands jeux de données, j'ai également pour objectif d'analyser les effets d'une combinaison de prédicteurs. Les effets cumulés de plusieurs facteurs environnementaux mettent en évidence des processus qui seraient indétectables avec une analyse univariée.

La croissance simultanée des données moléculaires et environnementales, en conjonction avec l'analyse multivariée, augmentent considérablement le nombre de modèles à traiter.

2.2 Changement de paradigme

Le séquençage du génome entier offre de nouvelles perspectives mais change radicalement les modalités d'analyse. Les projets utilisant des marqueurs génétiques peu onéreux peuvent analyser beaucoup d'individus. Le projet ECONOGENE a notamment étudié les races locales de moutons en Europe en échantillonnant 1'748 individus et 31 marqueurs microsatellites (soit 564 variantes possibles) (Peter et al., 2007). A contratio l'analyse des petits ruminants marocains dans le cadre de NextGen inclut 164 individus de chaque espèce pour 16 millions

2. https://www.23andme.com/
3. Illustrations : (a) Mehta, ? ; Dieterich et al., 2006 ; (b) Mehta, ? ; Hedberg, 2013.

de sites polymorphiques. Cette évolution est schématisée sur la fig. 2.1.

La nombre d'individus séquencés dépend des ressources financières disponibles. Le séquençage intégral fournit une image précise du génome d'un individu, mais coûte plus cher que l'extraction de marqueurs génétiques prédéfinis. Les études de la diversité génétique et de l'adaptation passent donc d'un contexte avec beaucoup d'individus et peu de marqueurs à un contexte avec peu d'individus et beaucoup de marqueurs.

2.3 Choix des individus

Le changement de paradigme et les contraintes financières limitent la taille de l'échantillon. Le faible nombre d'individus séquencés réduit la puissance des analyses statistiques. Simultanément, au vu du grand nombre de modèles à traiter, des tests de significativité conservateurs sont nécessaires pour éviter d'accumuler les faux positifs (erreurs de type I). Ces deux effets sont antagonistes : la sélection des modèles doit ainsi atteindre un équilibre pour détecter les signatures de sélection tout en rejetant les faux positifs.

En conséquence, le bon déroulement des analyses passe par un choix minutieux soit des individus à échantillonner, soit des échantillons à séquencer. Il s'agit d'optimiser la représentativité environnementale des échantillons afin de maximiser l'information fournie par les modèles. Les individus choisis doivent représenter toutes les races locales. Pour une raison technique, les échantillons doivent être séparés en deux groupes pour le séquençage. Le premier groupe sera séquencé une, voire deux fois si la probabilité d'avoir commis des erreurs de lecture est trop élevée. Ces échantillons doivent ainsi fournir une représentation fidèle des différents habitats. Si la qualité des données lors du premier séquençage est satisfaisante (c'est-à-dire si le taux de couverture du génome est suffisant pour la suite de l'analyse), le second groupe sera séquencé à son tour. C'est pourquoi il convient de répartir judicieusement les échantillons.

Lors de l'étude de l'adaptation d'un organisme à son environnement, le projet doit être organisé en considérant la taille des jeux de données moléculaires attendus, le nombre et le choix des individus à échantillonner et les FACTEURS DE CONFUSION (*confounding factors*) liés à la démographie.

2.4 Buts et objectifs

Mes travaux en génomique environnementale concernent le développement de méthodes de calcul haute performance intégrant des statistiques spatiales. Ces outils vont permettre de détecter rapidement et efficacement des signatures de sélection naturelle dans des jeux de données moléculaires comprenant jusqu'à 15 millions de marqueurs. L'intégration de l'analyse spatiale va enrichir l'interprétation des résultats afin de prendre en compte les inconvénients des méthodes corrélatives. Je présente ici l'approche que j'ai adoptée pour :

Détecter les régions adaptatives du génome par des méthodes corrélatives prévues pour de grands jeux de données moléculaires, puis analyser les comportements spatiaux locaux des marqueurs potentiellement soumis à la sélection et comparer les résultats avec ceux obtenus en génétique des populations.

Les objectifs détaillés sont les suivants :

1. Développer les algorithmes et les outils nécessaires pour la détection de régions génomiques soumises à la sélection dans de grands jeux de données issus du séquençage du génome entier :
   (a) en version de bureau pour les analyses rapides,
   (b) en version distribuée entre plusieurs ordinateurs pour du calcul haute performance ;
2. Développer un algorithme qui permette un traitement optimisé des modèles multivariés ;
3. Implémenter des méthodes de mesure de l'autocorrélation spatiale ;
4. Implémenter des techniques simples de traitement et de visualisation des résultats ;
5. Développer une stratégie d'échantillonnage des individus qui maximise la production potentielle de nouvelles connaissances en fonction des ressources financières disponibles ;
6. Rendre ces outils accessibles au plus grand nombre, en terme de facilité d'utilisation et de compatibilité.

La réalisation des objectifs permettra de répondre aux questions suivantes :

1. Dans un contexte de séquençage du génome entier (*whole-genome sequencing,WGS*), quelle est l'efficacité de l'approche corrélative adoptée par rapport à celle des principales méthodes en génomique environnementale ?
2. Dans quelle mesure l'utilisation de modèles multivariés permet-elle d'améliorer l'efficacité des approches corrélatives ?
3. Dans quelle mesure l'utilisation des statistiques spatiales permet-elle d'améliorer l'efficacité des approches corrélatives ?
4. Les résultats obtenus par les méthodes corrélatives sont-ils comparables à ceux obtenus avec des méthodes de GÉNOMIQUE DES POPULATIONS ?
5. Quels sont les apports de l'analyse spatiale dans l'élaboration de la stratégie d'échantillonnage ?
6. Peut-on rendre les méthodes corrélatives facilement utilisables par des chercheurs sans spécialisation en bioinformatique ?

2.5 Méthode de recherche

La détection de la sélection naturelle par l'étude des corrélations entre les fréquences de marqueurs génétiques et les caractéristiques des habitats est à la base de la génomique environnementale. Cette discipline permet d'étudier l'adaptation locale en partant de postulats indépendants de ceux de la génétique des populations. En particulier, l'approche utilisant des régressions logistiques qui est employée ici a déjà été appliquée avec succès dans de nombreuses études. C'est pourquoi le présent travail ne constitue pas une démonstration de faisabilité de la génomique environnementale et se concentre sur l'applicabilité de cette approche aux grands jeux de données moléculaires.

L'émergence du génotypage à haute densité et du séquençage intégral étant récente, la charge de calcul était jusqu'alors en adéquation avec les outils disponibles. Le projet NextGen est, à notre connaissance, le premier à étudier l'adaptation locale avec plusieurs millions de marqueurs génétiques. La charge de calcul prévue était trop importante pour les méthodes existantes, et c'est ainsi que nous avons décidé de développer un logiciel capable de traiter de grands jeux de données moléculaires. La conception de Samβada ainsi que les méthodes qui sous-tendent son fonctionnement sont au coeur de mes recherches présentées ici. La planification de NextGen prévoyait que le séquençage des ruminants marocains serait achevé suffisamment tôt pour que je puisse analyser ces données durant ma thèse. Or le déroulement du projet a finalement conduit à ce que les premières données disponibles soient les génotypes des bovins ougandais. Notre cas d'étude se concentre donc sur un panel de SNPs qui, bien qu'étant plus petit que les données attendues, est déjà suffisamment grand pour démontrer que Samβada est apte à analyser de très grands jeux de données moléculaires.

2.6 Organisation des chapitres

Dans le reste du document, le chapitre 3 résume le développement et les travaux effectués à ce jour en génomique environnementale. La campagne d'échantillonnage et la préparation des données sont présentées au chapitre 4. Le chapitre 5 décrit le développement de Samβada, le logiciel conçu pour analyser de grands jeux de données en génomique environnementale. Les principes mathématiques sous-tendant Samβada et les autres méthodes utilisées sont exposés au chapitre 6. Le chapitre 7 présente les analyses pour détecter les signatures de sélection chez *B. taurus* et *B. indicus* en Ouganda. La discussion est menée au chapitre 8. Finalement le chapitre 9 apporte la conclusion et présente quelques perspectives.

3 État de la recherche

Dès la découverte des principes biologiques de l'hérédité, la génétique des populations a étudié les forces à l'origine de l'évolution et leurs effets, dont les signatures de la sélection naturelle dans le patrimoine génétique. Depuis, la génétique environnementale s'est spécialisée dans l'analyse spatiale de la diversité génétique tandis que la génomique environnementale vise à détecter la sélection naturelle par la comparaison directe du génome des individus avec leur habitat.

3.1 Génétique des populations

A l'aube du XX^{e} s., les biologistes sont confrontés à deux aspects apparemment contradictoires de l'évolution. Charles Darwin, par ses minutieuses observations de la faune et la flore, avait décrit une évolution continue et graduelle des caractères morphologiques sous l'effet de la sélection naturelle (Darwin, 1859), tandis que Gregor Mendel avait postulé l'existence d'un support discret de l'hérédité grâce à ses recherches sur les variétés hybrides de pois, travaux restés méconnus jusqu'alors (Mendel, 1866). La modélisation mathématique des gènes et de leur transmission marque la naissance de la génétique, terme inventé en 1905 par William Bateson. En 1908, Godfrey H. Hardy et Wilhelm Weinberg découvrent séparément le principe qui porte leurs noms : en l'absence de forces évolutives les fréquences alléliques dans une population infinie sont invariantes d'une génération à l'autre et, si les individus choisissent leurs partenaires au hasard (panmixie), les fréquences génotypiques peuvent être calculées directement à partir des fréquences alléliques (Hardy, 1908 ; Weinberg, 1908). En étudiant les drosophiles, Thomas H. Morgan dresse la première carte des chromosomes (découverts par Walther Flemming, Flemming, 1882 ; Paweletz, 2001) porteurs des gènes, et explique leur rôle dans la transmission des phénotypes (Morgan et al., 1915). En postulant que chaque gène a un effet si petit sur la morphologie que ses variations peuvent être considérées comme continues, les premiers généticiens réunissent les visions de Mendel et de Darwin de l'évolution. Dès 1920, les travaux de Sewall G. Wright, Ronald A. Fisher et John B. S. Haldane fondent la génétique des populations pour laquelle la variation du patrimoine génétique est décrite en termes de mutations, de sélection, de migration et de dérive génétique.

Jusqu'au milieu du siècle, les expérimentateurs ne peuvent pas étudier les gènes directement et doivent baser leurs analyses sur des caractéristiques observables (comme la couleur des yeux chez les mouches). Les mutations neutres sont alors difficiles à étudier et l'opinion dominante considère que la plupart des loci sont soumis à la sélection et que seul un petit nombre d'entre eux sont polymorphiques. Les généticiens importent ou créent de nouveaux concepts mathématiques pour décrire l'évolution des fréquences alléliques dans une ou plusieurs populations. Wright crée les « surfaces de VALEURS SÉLECTIVES » (*fitness landscapes*) pour représenter la relation entre le génotype et le succès reproducteur (Wright, 1931 ; Wright, 1932) tandis que Gustave Malécot utilise des chaînes de Markov et développe les concepts d'*identité par descendance* (*identity by descent*) et de *coefficient de parenté* (*coefficient of kinship*) pour étudier la cosanguinité (Malécot, 1948) .

L'analyse moléculaire de la variabilité génétique commence dans les années 1960 avec l'électrophorèse sur gel. Cette technique permet d'étudier les protéines contenues dans une cellule, qui sont codées dans les gènes, en les soumettant à un champ électrique. La substitution d'un acide aminé crée une variation de charge et les allèles (variantes) d'une protéine migrent ainsi à des vitesses différentes, ce qui permet de les distinguer sur le gel. L'étude de la variabilité protéinique chez l'humain et la drosophile révèle que le nombre de loci polymorphiques est bien plus élevé qu'attendu (Harris, 1966 ; Hubby et Lewontin, 1966 ; Lewontin et Hubby, 1966).

Motoo Kimura postule alors que la plupart des variations alléliques sont neutres et que leurs effets sur la valeur sélective peuvent être négligés (Kimura, 1968). Lorsque qu'une mutation apparaît chez un individu, sa propagation ou sa disparition sont gouvernées par la dérive génétique, une des quatres forces évolutives, qui est due au tirage aléatoire des ALLÈLES transmis à chaque génération. Le modèle de Kimura est connu comme la *Théorie neutraliste de l'évolution* (*Neutral theory of molecular evolution*) qui sera complétée avec Tomoko Ohta pour former la *Theorie quasi-neutraliste* (*Nearly neutral theory,* Ohta, 1973). Une vive controverse s'ensuit entre les partisans du « sélectionnisme » et ceux du « neutralisme » (Crow, 2008). La théorie neutraliste a permis d'expliquer de nombreuses observations et a également contribué au développement de l'analyse coalescente et de la détection des loci sous sélection. Elle est un des principaux modèles de l'évolution utilisés aujourd'hui bien que d'autres approches considèrent que la sélection naturelle de mutations récurrentes a globalement plus d'influence que la dérive génétique (Gillespie, 2000 ; Gillespie, 2001).

La détection de loci soumis à la sélection se base couramment sur des comparaisons entre populations. L'indice de fixation F_{ST}, proposé par Wright (1949), mesure la part de la variance des fréquences alléliques qui est due à la divergence entre populations. F_{ST} peut être défini comme le rapport entre la variance des fréquences alléliques mesurées dans chaque population et la variance estimée en tirant aléatoirement des allèles dans la population totale. Les statistiques F, dont l'indice de fixation fait partie, permettent de décrire la relation entre le taux d'hétérozygotie observé et sa valeur attendue selon l'équilibre de Hardy-Weinberg dans une population fragmentée. Ces indices permettent d'étudier la distribution et l'évolution des fréquences alléliques dans des populations apparentées. Leurs valeurs peuvent être toutefois

difficiles à estimer car les échantillons analysés ne forment qu'un sous-ensemble de toutes les populations sous-jacentes et ces populations ne représentent elles-mêmes qu'une des nombreuses réalisations possibles du même processus évolutif (Holsinger et Weir, 2009).

Le test de Lewontin et Krakauer (1973) utilise l'indice de fixation pour détecter les loci soumis à la sélection. Il se base sur le principe que les effets de la migration, de la dérive génétique et de la consanguinité doivent être similaires sur l'ensemble du génome alors que la sélection naturelle doit agir sur certains loci en particulier. Les auteurs proposent une expression de F_{ST} ainsi qu'une estimation de sa distribution pour des loci neutres. Les loci dont la valeur de la statistique est peu probable sous la distribution neutre sont potentiellement soumis à la sélection naturelle. Ce test n'est plus utilisé aujourd'hui mais il a posé les bases de la détection de singularités (*outliers*). Outre l'étude des bases génétiques de l'adaptation, l'identification des loci singuliers permet de les écarter de certaines analyses. La reconstitution de l'histoire démographique d'une population est par exemple basée sur les loci neutres.

Les progrès technologiques en matière de génotypage et l'avènement du séquençage permettent d'étudier simultanément de nombreux loci ou régions du génome. Ceux-ci doivent beaucoup à l'invention de la réaction en chaîne par polymérase (*polymerase chain reaction, PCR*) qui permet de répliquer des fragments d'ADN (Saiki et al., 1988). Tirant parti de ces avancées, la génomique des populations (*population genomics*) utilise un échantillonnage de loci répartis sur tout le génome pour comprendre les influences respectives des forces évolutives sur la variabilité génétique (Luikart et al., 2003). Reprenant le concept de Lewontin, elle suppose que la démographie et l'histoire évolutive d'une population ont des effets similaires sur tous les loci d'un génome tandis que les loci soumis à la sélection naturelle se comportent souvent différemment et présentent ainsi des singularités dans la variabilité génétique. Les principes de la génomique des populations sont à la base de nombreuses méthodes pour identifier les loci soumis à la sélection naturelle.

L'amélioration des méthodes de détection de singularités s'est faite graduellement. La démocratisation de l'informatique a permis de simuler des systèmes de plus en plus complexes afin de calculer numériquement la valeur de paramètres n'ayant pas d'expression analytique. Les critiques du test de Lewontin et Krakauer portaient en partie sur leur estimation de la variance de F_{ST}. En 1996, Beaumont et Nichols proposent une méthode pour la calculer à partir d'une distribution simulée de la valeur de F_{ST} en fonction de l'hétérozygotie pour des loci neutres. Ils utilisent un modèle en îles et des simulations coalescentes pour déterminer les valeurs minimales et maximales de F_{ST} compatibles avec le modèle neutre pour chaque valeur possible de l'hétérozygotie. Les loci soumis à la sélection présentent des indices de fixation supérieurs ou inférieurs à ces bornes (Beaumont et Nichols, 1996). La sélection équilibrante, qui maintient la variabilité présente à un locus dans une population, est toutefois plus difficile à détecter que la sélection positive, qui favorise une nouvelle mutation à ce locus. Le modèle de populations sous-jacent ne correspond pas forcément à la démographique réelle étudiée, ce qui peut provoquer l'apparition de faux positifs, mais la méthode est globalement robuste aux perturbations (Excoffier et Heckel, 2006). Plusieurs logiciels utilisent cette approche pour

la détection de signatures de sélection (`LOSITAN`, basé sur `fdist`, Antao et al., 2008 ; `Mcheza`, basé sur `DFDIST`, pour les marqueurs dominants, Antao et Beaumont, 2011). Excoffier et al. (2009) étendent ce modèle au cas d'une structure hiérarchique de populations (inclus dans `Arlequin`).

Vitalis et al. (2001) proposent un modèle où une population ancestrale a subi un goulot d'étranglement puis s'est séparée en deux populations qui ont évolué sans migrations sous l'effet de la dérive génétique. Les auteurs présentent une statistique de différentiation intra-population basée sur les observations. Cette statistique est reliée au temps depuis la divergence des populations. La distribution conjointe de cette statistique pour les deux populations est construite avec des simulations coalescentes pour diverses valeurs des paramètres du modèle (taux de mutation, durée du goulot d'étranglement). Les loci présentant des valeurs singulières des deux statistiques intra-populations par rapport aux valeurs simulées sont potentiellement soumis à la sélection. Si plus de deux populations sont étudiées, les comparaisons sont faites par paires de populations. Dans ce cas, les loci détectés dans plusieurs comparaisons sont des candidats pour l'adaptation. Le logiciel `detSel` permet de détecter les signatures de la sélection naturelle avec cette méthode (Vitalis et al., 2003).

L'approche adoptée par Foll et Gaggiotti (2008) permet de déterminer la probabilité qu'un locus soit soumis à la sélection . Ils modélisent F_{ST} avec deux types de paramètres de simulation représentant les effets propres à chaque population et les effects spécifiques à chaque locus. Une simulation bayésienne permet d'estimer la probabilité d'obtenir les fréquences alléliques observées en fonction de ces paramètres. La sélection peut être ajoutée ou enlevée du modèle pour chaque locus à chaque itération. La probabilité qu'un locus soit soumis à la sélection correspond à la fraction des itérations incluant le paramètre correspondant. Foll et Gaggiotti développent le logiciel `BayeScan` qui a popularisé leur approche.

D'autres statistiques sont également utilisées pour détecter les effets de la sélection (Nielsen et al., 2005 ; Volis, 2008). Lorsqu'une mutation bénéfique apparaît, elle se répand dans la population, éventuellement jusqu'à sa fixation. Les mutations neutres situées sur des loci proches vont également se propager sous l'effet du déséquilibre de liaison. Ces phénomènes sont connus sous les noms de balayage sélectif (*selective sweep*) et d'auto-stop génétique (*genetic hitchhicking*). Leur conséquence est de modifier la distribution des fréquences des mutations autour des loci soumis à la sélection par rapport au reste du génome. Plusieurs statistiques, comme le *D* de Tajima (1989), mesurent les déformations de ce spectre de fréquences (*site frequency spectrum*) afin de détecter des signatures de sélection. D'autres méthodes se basent sur les variations du déséquilibre de liaison pour identifier les loci adaptatifs (Sabeti et al., 2002 ; voir aussi Nielsen et al., 2005 ; Volis, 2008). Plusieurs démarches analysent quant à elles la fréquence des mutations polymorphiques intraspécifiques (à l'intérieur d'une espèce) et interspécifiques (entre plusieurs espèces). Les gènes présentant des taux de polymorphismes singuliers sont susceptibles d'être soumis à la sélection naturelle (test de Hudson-Kreitman-Aguade, Hudson et al., 1987). Ces comparaisons peuvent être combinées avec le décompte des mutations synonymes et non-synonymes lors de la transcription des régions codantes

du génome. Les mutations synonymes ne modifient pas la séquence d'acides aminés (qui forment les protéines) et sont généralement considérées comme neutres, alors que les mutations non-synonymes peuvent avoir un effet phénotypique et peuvent ainsi être sujettes à la sélection naturelle. Le test de MacDonald-Kreitman compare ainsi les taux de mutations synonymes et non-synonymes entre deux espèces proches. Les loci présentant des valeurs singulières sont des candidats potentiels pour l'adaptation (McDonald et Kreitman, 1991).

3.2 Génétique spatiale et approches corrélatives

La génétique des populations a rapidement pris en compte les effets de la séparation géographique entre groupes d'individus. Wright (1931) étudie comment la dérive génétique et les migrations influencent les fréquences alléliques de populations disjointes, ce qui le conduira à concevoir le modèle en îles. Les populations ayant une distribution continue sur le territoire peuvent également diverger. Si leur distribution spatiale est beaucoup plus grande que la distance que parcourent les individus pour se reproduire, le FLUX DE GÈNES dû à la MIGRATION ne compense pas la dérive génétique. Les fréquences alléliques des groupes géographiquement éloignés peuvent alors se différencier au fil des générations, c'est l'isolation par la distance (Wright, 1943). Le modèle en îles est généralement utilisé pour formuler l'hypothèse nulle lors d'un test de neutralité. Or il implique que toutes les populations peuvent échanger des migrants, ce qui n'est pas le cas pour les populations réelles soumises à l'isolation par la distance. Meirmans (2012) expose comment les tests de neutralité basés sur des différences de fréquences alléliques et les approches corrélatives peuvent présenter de nombreux faux positifs en cas d'isolation par la distance entre les populations. L'auteur explique que l'autocorrélation spatiale est inhérente à tout processus biologique. A moins que les fréquences alléliques dans les populations étudiées soient indépendantes de la distance les séparant, l'espace ne peut pas être considéré comme statistiquement neutre. L'auteur invite les chercheurs à intégrer des modèles nuls tenant compte de l'isolation par la distance dans leurs analyses.

La découverte de l'isolation par la distance a rapidement incité les chercheurs à considérer l'influence de l'espace géographique sur la différenciation entre populations. Dès 1945, Ernest Mourant a étudié les différents systèmes de groupes sanguins et leur capacité à expliquer les incompatibilités observées lors de transfusions. Il a intégré l'effet de la distance géographique dans ses recherches par la cartographie et la comparaison de la distribution des groupes sanguins présents dans de nombreuses populations humaines (p. ex. Chalmers et al., 1953). L'étude de la synthèse de la vitamine D lui a également permis de mettre en évidence l'influence de l'environnement, ici de la quantité d'ensoleillement reçue, dans la régulation de processus biologiques (Mourant et al., 1976).

L'isolation par la distance dépend de l'équilibre entre mutations et migrations et peut être analysée avec la distance de dispersion des individus et la densité de population. Luigi Luca Cavalli-Sforza réunit les modèles développés par Wright avec les marqueurs génétiques disponibles chez l'humain et étudie la distribution spatiale des variations de fréquences alléliques

en utilisant la densité d'habitations et le taux de consanguinité pour mesurer d'isolation par la distance (Cavalli-Sforza, 1966). Il démontre ainsi que la distribution des groupes sanguins est homogène dans les villes alors qu'elle peut varier entre des villages voisins en montagne dans des proportions dépassant les effets aléatoires dus à l'échantillonnage. L'auteur et Anthony W. F. Edwards utilisent ensuite cinq loci échantillonnés dans 15 populations réparties dans le monde pour tenter de reconstituer l'évolution des populations humaines. Ils considèrent tous les arbres phylogénétiques possibles et sélectionnent celui qui explique le mieux la distribution de fréquences alléliques observée grâce à une méthode basée sur le maximum de vraisemblance.

La génétique des populations assimile des méthodes d'analyse spatiale avec les travaux de Sokal et Oden sur l'autocorrélation spatiale (Sokal et Oden, 1978[a] ; Sokal et Oden, 1978[b]). Ils développent l'usage de corrélogrammes pour analyser la dépendance spatiale des fréquences alléliques. Les auteurs proposent une méthode basée sur la comparaison des distances géographiques et génétiques pour identifier les processus évolutifs influençant les fréquences alléliques dans chaque population. L'habitat des individus n'est pas encore pris en compte dans ces modèles.

Les premières études de l'influence de l'environnement sur le polymorphisme génétique sont menées par Jeffry B. Mitton chez le pin jaune (*Pinus ponderosa*). Deux analyses par électrophorèse révèlent une variation de l'activité enzymatique suivant l'orientation du terrain et l'altitude, qui modifient l'humidité des habitats (Mitton et al., 1977 ; Mitton et al., 1980). L'épinette d'Engelmann (*Picea engelmannii*) présente également une variation de son activité enzymatique en fonction de l'humidité du sol (Stutz et Mitton, 1988). Les populations étudiées conservent des différences génétiques malgré leur proximité et des échanges avérés de gènes, l'effet de la sélection naturelle est ici plus important que celui des migrations. Ce type d'analyse peut aussi s'appliquer à la pression de sélection exercée par les insectes herbivores. Dans une population de *Pinus edulis* soumise à une intense prédation de phalènes (*Dioryctria albovitella*), Mopper et al. (1991) démontrent que les arbres résistants aux attaques des insectes sont plus souvent hétérozygotes que les autres pour deux loci impliqués dans la défense parasitaire et produisent ainsi plus de résine pour cicatriser leurs plaies.

Ces études de la diversité génétique des conifères en fonction de leur habitat ouvrent la voie aux approches corrélatives utilisées aujourd'hui.

3.3 Génomique environnementale

L'étude de la distribution spatiale de la variabilité génétique en relation avec l'environnement est appelée GÉNÉTIQUE ENVIRONNEMENTALE à partir de la publication de l'article fondateur de Manel et al. en 2003. Ce domaine est à l'intersection de l'écologie environnementale, de la génétique des populations et de l'analyse spatiale et s'intéresse à l'influence de l'environnement sur les processus évolutifs, en particulier les flux de gènes et l'adaptation locale. Les études menées dans cette discipline consistent souvent à détecter des discontinuités dans la distribu-

tion spatiale des marqueurs génétiques et à les relier avec des éléments topographiques ou environnementaux (Coulon et al., 2006 ; Ortego et al., 2012).

Les études en génétique environnementale doivent si possible utiliser l'individu comme unité d'analyse (Manel et al., 2003). Les populations naturelles étant rarement distribuées en clusters distincts, cette approche permet d'étudier des processus locaux sans délimiter des populations au préalable. La stratégie d'échantillonnage est essentielle et doit couvrir tout le territoire étudié sur la base d'un plan stratifié, afin de représenter les habitats où vit l'organisme dans cette région (Manel et al., 2010). La génétique environnementale permet aujourd'hui d'étudier les flux de gènes dans des environnements hétérogènes et fragmentés et permet d'estimer la connectivité fonctionnelle (les déplacements des individus) entre les habitats (Manel et Holderegger, 2013).

La génétique environnementale et la génomique des populations sont apparues simultanément. Dans les références commentées de leur article, Luikart et al. (2003) mentionnent déjà l'idée de les combiner afin d'étudier des associations directes entre le génome et l'environnement, la *génomique environnementale*. C'est ainsi que Joost et al. (2007) proposent une nouvelle méthode utilisant un *modèle linéaire généalisé* (*generalized linear model, GLM)* pour détecter les signatures de sélection. Leur approche, *spatial analysis method (SAM)*, est basée sur le concept de *coïncidence spatiale* : Les échantillons sont géo-référencés et leurs coordonnées servent à caractériser leurs habitats au moyen de bases de données environnementales (topographiques et climatiques). Une fois génotypé, chaque individu est décrit par des marqueurs génétiques et des variables environnementales. Les associations sont modélisées par une régression logistique pour chaque paire {marqueur, variable}. La significativité de ces associations est analysée avec les tests du rapport des vraisemblances et de Wald. L'hypothèse nulle est que le modèle incluant la variable environnementale n'explique pas la distribution du marqueur plus précisément qu'un modèle n'utilisant qu'une constante (la fréquence moyenne du marqueur). Les auteurs étudient un groupe de 367 grands charançons du pin (*Hylobius abietis*) collectés dans 20 sites à travers l'Europe et génotypés avec 83 marqueurs AFLP. En utilisant 10 variables environnementales pour décrire les habitats, 11 loci sont détectés comme étant potentiellement soumis à la sélection. Les auteurs comparent leurs résultats avec ceux obtenus avec la méthode de Beaumont et Nichols (1996). Leur approche détecte légèrement plus de marqueurs, mais les découvertes concordent. Dans la foulée, Joost, Kalbermatten et Bonin (2008) présentent `MatSAM`, probablement le premier logiciel dédié à la génomique environnementale basé sur une approche corrélative. `MatSAM` utilise une librairie `MATLAB` pour calculer les régressions logistiques et fournit les résultats sous forme de matrices qui peuvent être importées dans un tableur. Le paquet inclut des macros permettant de choisir le seuil de significativité et d'appliquer la correction de Bonferroni dans `Excel`. `MatSAM 2` permet en plus d'inclure des variables environnementales qualitatives, ordinales et nominales, dans les modèles (Joost et al., 2012).

D'autres méthodes ont été proposées pour modéliser la distribution spatiale d'un allèle en fonction de facteurs environnementaux. Certaines d'entre elles, comme les *generalized es-*

timating equations (*GEE*, Carl et Kuehn, 2007), ciblent l'analyse de données contenant des pseudoréplicats. Ce sont des échantillons qui ont été récoltés au même endroit et sont donc décrits par les mêmes variables environnementales. Or leur similitude pourrait être due à l'autocorrélation spatiale (par exemple s'ils sont apparentés) plutôt qu'au fait qu'ils partagent le même habitat. Les GEE sont un prolongement des GLM tenant en partie compte de l'autocorrélation spatiale. Une matrice de corrélation sert à représenter les similitudes entre les échantillons qui ont été récoltés au même endroit. Poncet et al. (2010) utilisent des GEE pour étudier l'adaptation locale de deux populations d'*Arabis alpina* avec 825 marqueurs AFLP. Les auteurs relèvent que les GEE ne sont pas basés sur la vraisemblance et ne permettent donc pas de calculer certaines statistiques couramment utilisées (G, AIC). Comme ces modèles n'intègrent pas l'autocorrélation spatiale entre les lieux de récolte, l'isolation par la distance entre les individus peut être un facteur de confusion. Les auteurs ont testé l'indépendance de leurs deux populations et ont trouvé 4 loci détectés en commun (sur 78). Cette stratégie d'échantilllonnage réduit le risque de faire une fausse découverte due à la structure de population.

Un autre perfectionnement des GLM utilise des effets aléatoires pour intégrer les similarités fortuites entre individus proches. Un modèle linéaire généralisé mixte (*generalized linear mixed model, GLMM*) est un modèle linéaire généralisé hiérarchique qui tient compte de la dépendance entre certaines observations (Bolker et al., 2009). Les GLMM permettent de décrire de nombreux phénomènes mais requièrent des méthodes d'estimation plus perfectionnées que les GLM. Santos-del-Blanco et al. (2012) étudient la relation entre la taille des pins martimes (*Pinus pinaster*) au moment de la première floraison et le genre des fleurs produites. (Ces arbres peuvent porter des fleurs mâles et femelles.) Les auteurs ont planté 2'767 pins en provenance de 5 régions d'Europe dans une pépinière. Ils ont mesuré la taille des arbres sur plusieurs années et l'ont modélisée en fonction de la provenance, du type de fleurs produites et de la famille d'origine. Ce dernier paramètre est inclus sous forme d'un effet aléatoire dans le modèle mixte afin de tenir compte des pseudoréplicats. La plupart des arbres ont une taille similaire lorsqu'ils commencent à produire des fleurs mâles, alors que leur taille varie suivant les régions d'origine au moment de la pousse des premières fleurs femelles. Les auteurs relèvent que ce deuxième phénomène est lié à la rudesse de l'environnement (comme la température moyenne du mois le plus froid) et que cette association reste significative en tenant compte des populations d'origine. En effet, la production de fleurs femelles requiert de l'énergie et retarde la croissance : les pins maritimes habitués à un climat hostile en produiraient rapidement tandis que les arbres provenant de régions plus clémentes retarderaient leur pousse pour croître et concurrencer leurs voisins.

La structure spatiale des points d'échantillonnage peut être utilisée pour tenir compte de l'effet de variables environnementales inconnues (Manel et al., 2010). Les *Moran eigenvector map (MEM)* sont une approche possible pour identifier la structure spatiale à différentes échelles (Borcard et Legendre, 2002 ; Dray et al., 2006). Elles sont basées sur la diagonalisation d'une matrice de pondération spatiale qui représente la similitude (ou la distance) entre les points. Les vecteurs propres obtenus décrivent la structure spatiale de l'échantillon à différentes

échelles. Les *principal coordinates of neighour matrices (PCNM)* sont un cas particulier des MEM où la matrice de pondération est basée sur la distance et comporte deux valeurs : un poids pour toutes les distances inférieures ou égales à un seuil et un autre poids (plus grand) pour les distances supérieures (Dray et al., 2006). Manel et al. (2010) étudient l'adapation locale de deux populations d'*A. alpina* dans les Alpes aux moyen de huit variables environnementales et de neuf MEM à large échelle. Les auteurs détectent différents loci potentiellement soumis à la sélection suivant l'échelle géographique de l'analyse. L'inclusion des MEM leur permet de détecter des loci dont la structure spatiale varie à une échelle qui n'est pas représentée par les variables environnementales utilisées. Bothwell et al. (2013) comparent une approche corrélative basée sur les PCNM avec deux approches de détection de singularités (Beaumont et Balding, 2004 ; Foll et Gaggiotti, 2008). Leur étude porte sur 218 *Gentiana nivalis* génotypées pour 157 marqueurs AFLP. Leur approche corrélative détecte plus de loci potentiellement soumis à la sélection que les méthodes basées sur les loci singuliers (15 contre 9). Les auteurs attribuent ces détections à des variations locales dans la distribution des marqueurs. Ces variations ne sont pas détectables à l'échelle des quatre populations de *G. nivalis* utilisées par les deux autres approches. Ces deux études transforment les variables environnementales avec des fonctions cubiques pour incorporer des effets non-linéaires en plus des variables d'origine dans les modèles.

Les *generalized additive models (GAM)* incluent une transformation de certaines variables explicatives dans le but de décrire des relations non-linéaires avec les observations (Hastie et Tibshirani, 1986). Les *generalized additive mixed models (GAMM)* sont un prolongement des GAM et des GLMM où la transformation est une fonction lisse (infiniment dérivable) (Lin et Zhang, 1999). Benson et al. (2012) utilisent des GAMM pour leur étude des croisements entre loups de l'Est (*Canis lycaon*), loups gris (*C. lupus*) et coyotes (*C. latrans*) en Ontario (Canada). Les loups de l'Est sont une espèce protégée au Canada et s'hybrident souvent avec les coyotes et les loups gris. Les auteurs ont échantillonné 342 individus à l'intérieur et aux alentours du Parc provincial Algonquin où vit une population distincte de loups de l'Est. La fraction d'appartenance de chaque individu aux trois populations a été mesurée avec 12 marqueurs microsatellites. Le modèle expliquant la distribution spatiale des génotypes incluait la distance depuis le centre du parc, la densité de cerfs et d'élans (ces derniers sont principalement chassés par les loups) et l'interaction humaine (densité de routes). La proportion d'ascendance loup gris de chaque individu était modélisée par une fonction lisse de la distance au centre du parc, de la densité de proies et de la densité de routes (partie additive du modèle). Chaque individu était également caractérisé par sa meute (effets aléatoires du modèle mixte). Les résultats montrent que les coyotes ne se sont pas introduits dans le Parc car les cerfs y sont rares, particulièrement en hiver. A l'extérieur du Parc, les croisements forment des motifs spatiaux irréguliers, certains individus étant clairement issus d'une population. Les loups de l'Est y vivent principalement dans les zones éloignées des routes où les élans sont nombreux. Le Parc forme donc un habitat idéal pour cette espèce. Benson et al. (2012) concluent que les loups de l'Est sont restés une population distincte dans le Parc car les coyotes ne s'y aventurent pas. D'autres analyses sont nécessaires pour déterminer si les croisements observés à l'extérieur

du Parc sont dus à des préférences de reproduction ou aux facteurs environnementaux (routes et élans).

Comme mentionné précédemment, la distribution spatiale d'un marqueur génétique peut être due à l'adaptation des individus aux conditions locales sous l'effet de la sélection naturelle, ou à l'histoire démographique de la population - comme une dispersion limitée des individus ou la rencontre de deux groupes jusqu'alors séparés. Plusieurs approches visent à démêler les variations de fréquences alléliques d'origine adaptatives ou démographiques. Notamment, Coop et al. (2010) proposent une solution en deux étapes, `BayEnv`, qui s'appuie sur l'analyse bayésienne : Un ensemble de loci neutres sert d'abord à estimer les corrélations des fréquences alléliques entre les populations (qui sont supposées être issues de la même population ancestrale). La deuxième partie du test utilise tous les loci ainsi que les variables environnementales. Les corrélations entre les populations permettent de modifier l'hypothèse nulle du test (aucune différence de fréquences entre les populations) pour tenir compte de l'histoire démographique (différences de fréquences dues à la dérive génétique). Coop et al. relèvent que leur approche s'apparente à un GLMM. `BayEnv` requiert l'identification préalable d'un sous-ensemble de loci neutres ainsi que l'assignation de chaque individu à une population. Lorsque les individus étudiés sont distribués de manière continue dans l'espace, la détermination des populations passe par une analyse basée sur les loci neutres. Les auteurs ont analysé un jeu de 640k SNPs collectés chez 927 humains issus de 52 populations réparties dans le monde (*human genome diversity project*, Li et al., 2008). La matrice de covariance révèle la similitude entre populations géographiquement proches. Les variables environnementales sont la latitude et les précipitations en été. Les modèles ayant les plus hauts facteurs de Bayes correspondent à des loci impliqués dans le fonctionnement du système immunitaire, la pigmentation de la peau et la couleur des yeux. Cette approche a depuis été améliorée en incluant la possibilité de « corriger » les fréquences alléliques pour supprimer les effets de l'échantillonnage et de la structure de population. `BayEnv2` permet également de traiter les données produites avec des techniques de séquençage groupé (où plusieurs échantillons sont analysés en même temps) (Günther et Coop, 2013).

Toujours dans le but de prendre en compte la structure de population, Frichot et al. (2013) proposent de compléter le modèle linéaire en y ajoutant un terme pour tenir compte des phénomènes non observés. Ce terme représente l'effet de variables cachées, il s'apparente à une estimation de la matrice de corrélations entre les populations et donne son nom à la méthode : *latent factor mixed model (LFMM)*. L'estimation de ces paramètres supplémentaires permet de mesurer les variations de fréquences alléliques non-expliquées par l'environnement et qui sont ainsi attribuées à la démographie. LFMM s'inspire de méthodes probabilistes pour l'analyse en composantes principales et la factorisation de matrices (Tipping et Bishop, 1999 ; Salakhutdinov et Mnih, 2008). L'estimation des paramètres se déroule également dans un cadre bayésien. Les auteurs comparent les résultats de LFMM à ceux d'autres approches pour la détection de loci soumis à la sélection, dont `MatSAM` et `BayEnv`. Ils indiquent que LFMM détecte beaucoup moins de faux positifs que le premier s'il y a une structure de population et qu'il est plus rapide que le second.

La structure de population peut aussi être incluse comme un paramètre spatial du modèle. Guillot et al. (2013) proposent une modification du modèle logistique nommée *spatial generalized linear mixed model (SGLMM)*. Le modèle inclut un terme aléatoire représentant l'autocorrélation spatiale due à l'histoire de la population. La densité de probabilité sous-jacente est la même pour tous les loci. La méthode suppose ainsi que les variations des fréquences alléliques non décrites par les variables environnementales agissent à une échelle géographique particulière, et qu'elles sont produites par la structure spatiale de la population étudiée. La technique utilisée pour ajuster le modèle est spécifique au type de terme qu'ils ajoutent et évite ainsi de recourir à une estimation bayésienne basée sur l'algorithme de Monte-Carlo (INLA Rue et al., 2009 ; Lindgren et al., 2011). Les auteurs comparent leur approche à celle de Joost, Bonin, Taberlet et al. (2008) en analysant les charançons du pins avec un sous-ensemble de quatre variables environnementales. SGLMM détecte 3 loci associés à la même variable environnementale alors que SAM en détecte 11 répartis entre les variables. Les loci identifiés par SAM avec les associations les plus significatives sont ceux détectés par SGLMM.

Les méthodes de détection de signatures de la sélection naturelle se sont développées et complexifiées grâce aux progrès réalisés en matière de génotypage et de séquençage mais aussi grâce à l'accroissement des moyens de calcul à disposition des chercheurs et à la sophistication graduelle des méthodes statistiques. La validation des résultats obtenus nécessite d'étudier les fonctions biologiques liées aux loci identifiés, or ce travail est colossal. Fort heureusement, l'avènement des ordinateurs a aussi ouvert la voie aux simulations numériques (Epperson et al., 2010). De nombreux modèles informatiques décrivent l'évolution d'une population au cours du temps. Une simulation permet de contrôler les pressions évolutives auxquelles les individus sont soumis, de simuler l'évolution du patrimoine génétique sur un grand nombre de générations et de répéter plusieurs fois l'expérience pour comparer les populations obtenues. Les simulations numériques permettent également de tester les prédictions des méthodes de détection de singularités et de comparer leurs résultats dans un cadre où les loci soumis à la sélection sont connus. La plupart des modèles récents sont basés sur les individus et leur permettent de migrer, le patrimoine génétique étant transmis aux descendants selon l'hérédité mendélienne. Par exemple, `quantiNEMO` (Neuenschwander et al., 2008) simule des individus répartis en populations sans tenir compte de leur position (modèle en îles ou en pierres de gué (*stepping stones*)). Chaque individu peut avoir plusieurs caractéristiques phénotypiques liées à plusieurs loci, il peut aussi posséder des loci neutres. La valeur adaptative d'un phénotype dépend de la population dans laquelle l'individu se trouve et peut varier au cours du temps. `CDPOP` (Landguth et Cushman, 2010) modélise le paysage par un réseau de positions géographiques possibles et des coûts de déplacement. Les générations peuvent se chevaucher et le paysage peut être modifié durant la simulation. La dernière version permet d'ajouter un ou deux loci soumis à la sélection. La valeur adaptative des individus est déterminée par une surface décrivant la mortalité à la naissance (Landguth et al., 2012). Finalement, `SimAdapt` (Rebaudo et al., 2013) utilise une grille régulière où plusieurs individus peuvent occuper la même cellule. Le nombre de loci soumis à la sélection n'est pas limité et la

valeur adaptative dépend du type de terrain dans la cellule. Le paysage peut varier au cours du temps afin d'étudier les effets d'un changement d'occupation du sol ou d'une fragmentation du territoire sur la population.

Voilà à ce jour la composition du monde des approches corrélatives au sein duquel les développements que je propose viennent s'inscrire. Le chapitre qui suit décrit les données utilisées, explique comment l'information génétique a été produite et propose une approche qui permet de sélectionner les échantillons potentiellement les plus informatifs dans le cadre de l'application d'une approche corrélative.

4 Échantillonnage et données

Les échantillons récoltés sur le terrain doivent fournir une image fiable de leur population d'origine, en matière de diversité génétique et de variété des habitats. Lorsque la campagne doit être menée d'une traite sur un grand territoire, l'optimisation de la représentativité spatiale des lieux de récolte permet d'éviter les biais d'un échantillonnage improvisé. La connaissance des caractéristiques environnementales des habitats permet ensuite de sélectionner les individus qui seront séquencés afin de maximiser l'information disponible pour les analyses.

4.1 Échantillonnage

4.1.1 Stratégie

La collecte des données est une étape essentielle de toute recherche scientifique. Le bon déroulement des analyses et la validité des conclusions qui peuvent en être tirées dépendent de la stratégie d'échantillonnage adoptée. Lors d'une étude en laboratoire, les paramètres de l'expérience peuvent être choisis indépendamment les uns des autres. L'effet de leur interaction peut alors être mesuré et son influence sur les résultats peut être minimisée en suivant un plan d'expérience approprié (*Design of experiments,* Box et al., 1978). Cette approche n'est pas directement applicable en écologie car les paramètres de l'expérience sont déterminés par l'environnement. Or les caractéristiques d'un milieu sont souvent interdépendantes et les habitats proches ont tendance à se ressembler (autocorrélation spatiale, voir sec. 6.1.2). Des organismes voisins ont aussi plus de chances de se ressembler que des organismes distants, soit parce qu'ils partagent le même habitat, soit parce qu'ils sont apparentés. Les échantillons doivent donc être récoltés selon un schéma prédéfini dans des habitats distincts qui représentent de manière fiable la variété de l'environnement.

Outre les considérations statistiques, un échantillonnage improvisé peut également être biaisé par le comportement de l'expérimentateur qui pourrait être tenté de visiter les lieux les plus faciles d'accès ou qu'il connaît le mieux (Guisan et Zimmermann, 2000). Dans le cadre

d'une étude sur les animaux domestiques, les éleveurs réunis en associations sont également susceptibles de posséder des animaux apparentés. Les registres généalogiques (*herdbooks*), s'ils existent, permettent de les identifier.

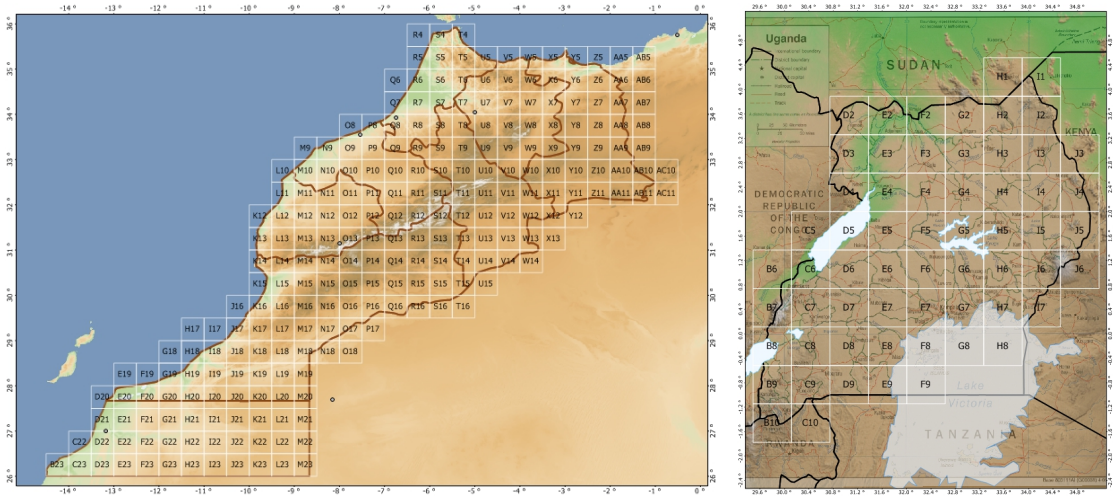
Une campagne d'échantillonnage doit donc prendre en compte l'espace géographique où vit l'organisme en question, l'espace écologique (ou environnemental) qui caractérise son habitat et l'espace biologique des caractères étudiés. Les schémas de récolte basés sur un échantillonnage dans l'espace des variables environnementales ou des caractères biologiques fournissent en général des estimations plus précises qu'un échantillonnage dans l'espace géographique (Manel et al., 2012). Les études sur la faune et la flore sauvages utilisent souvent un échantillonnage aléatoire où les lieux de récolte sont tirés au sort. Cette méthode limite les effets de l'autocorrélation spatiale et les biais dus à l'expérimentateur. D'autres approches sont basées sur une classification de l'espace environnemental et un échantillonnage à l'intérieur de chaque strate. Elles permettent de décrire précisément l'influence des facteurs environnementaux sur les caractères biologiques. Les approches dérivées de la méthode des surfaces de réponse (Box et al., 1978, chap. 15) détectent les extrema ou les crêtes du caractère biologique dans l'espace des paramètres environnementaux. Ces méthodes sont cependant difficilement applicables à une étude sur les animaux d'élevage à l'échelle d'un pays. Les coordonnées géographiques qui seraient tirées aléatoirement ou choisies à partir de l'espace des paramètres auraient peu de chances de désigner l'emplacement — même approximatif — d'une ferme. En particulier quand les informations disponibles dans les pays étudiés ne permettent pas de déterminer à l'avance où se situent lesdites fermes. De plus, la méthode des surfaces de réponse nécessite plusieurs échantillonnages et analyses successifs pour explorer la distribution des données biologiques, ce qui est exclu dans le contexte de NextGen pour des raisons financières.

4.1.2 Déroulement de la campagne

La collecte des données du projet NextGen a été organisée pour assurer une distribution spatiale représentative de toutes les régions du Maroc et de l'Ouganda. Comme l'emplacement des fermes n'était pas connu à l'avance, nous avons utilisé une grille régulière pour couvrir aussi uniformément que possible les territoires. Les échantillonneurs gardaient ainsi une marge de manoeuvre pour parer aux aléas du terrain. Les grilles utilisées pour quadriller les pays sont présentées sur la fig. 4.1.

Au Maroc, l'étude concerne l'adaptation des petits ruminants (chèvres et moutons) à leur environnement local. Les équipes d'échantillonnage ont visité trois fermes par cellule et y ont prélevé des échantillons de peau sur trois animaux de chaque espèce. Ils ont également relevé la race et des caractères phénotypiques comme l'âge, la longueur des cornes ou la couleur du pelage. Cela concernait environ 18 animaux par cellule, soit 1'283 chèvres et 1'412 moutons pour les 164 cellules du pays.

En Ouganda, quatre fermes ont été visitées dans chaque cellule et quatre bovins, ankoles ou



(a) Maroc (b) Ouganda

Figure 4.1 – Grilles utilisées pour l'échantillonnage au Maroc et en Ouganda. La taille des cellules a été ajustée en fonction des conditions sur le terrain, de la résolution spatiale des données environnementales disponibles (Maroc) et de la prévalence des cas d'individus touchés par la trypanosomiase (Ouganda). Les cellules mesurent 0,5° × 0,5° au Maroc et 0,63° × 0,63° en Ouganda.

zébus, y ont été échantillonnés. Les données récoltées comprenaient un morceau de peau, des données morphologiques et une prise de sang, afin de pouvoir détecter les parasites présents. Un total de 917 bovins ont été échantillonnés dans les 51 cellules en Ouganda.

En raison d'impératifs financiers, les analyses génétiques ont dû être organisées en fonction des types d'investigation. L'étude de l'adaptation à l'environnement était prévue sur le génome entier des chèvres et des moutons. Or le coût de ces analyses au printemps 2012 permettait de séquencer 164 individus de chaque espèce au Maroc. La section 4.4 présente la méthode utilisée pour choisir les échantillons à séquencer en tenant compte des contraintes techniques et en optimisant la représentativité de l'environnement.

En revanche, tous les bovins ougandais ont été génotypés pour étudier leur résistance aux parasites. Environ cent d'entre eux l'ont été avec une puce à haute densité (800k SNPs) et les quelques huit cents autres ont été génotypés avec une puce à densité moyenne (54k SNPs). En outre et à l'inverse du Maroc, les conditions climatiques sont relativement uniformes en Ouganda. Les partenaires de NextGen à l'Institut de Zootechnique de l'Université Catholique du Sacré Coeur à Piacenza (en Italie) ont donc choisi deux individus dans chaque cellule pour le génotypage à 800k, les autres individus étant génotypés à 54k. Cette technique étant plus rapide que le séquençage intégral, les données des bovins sont les seules disponibles et utilisables dans le cadre de la présente thèse. Elles sont décrites à la section 4.5 et analysées au chapitre 7.

4.1.3 Supervision

La campagne d'échantillonnage est une composante essentielle du projet NextGen. Elle s'est échelonnée sur plus d'une année et a mobilisé plusieurs équipes dans chaque pays. Son organisation et son bon déroulement nécessitaient un outil informatique de supervision permettant de :

— collecter et centraliser les données manuscrites sur le terrain ;

— repérer et corriger les erreurs de saisie ;

— suivre état d'avancement de l'échantillonnage.

Nous avons décidé de créer un site Internet couplé à une base de données. Les partenaires ont ainsi pu facilement enregistrer, modifier et visualiser leurs données et la centralisation des informations a facilité leur sauvegarde. J'ai utilisé `EasyDev` pour développer le site (Mingard, 2008). A partir d'un script décrivant les tables et leur relations, `EasyDev` génère la base de données `MySQL` et une interface pour y accéder en `PHP`. Il permet aussi d'ajouter des méthodes de recherche spécifiques.

L'unité d'échantillonnage est l'individu et la structure de la base de données reflète cette organisation. Les principales caractéristiques d'un animal sont sa position (en coordonnées sphériques), son espèce, sa race et son numéro. Son profil dans la base de données peut être complété avec une photo et un lien vers le profil de l'éleveur. Chaque échantillon s'est vu attribuer un identifiant unique basé sur le pays, l'espèce, la cellule et le numéro. Lors de la saisie de plusieurs individus, certaines informations comme la position ou la localité la plus proche peuvent être reprises de l'enregistrement précédent pour éviter les erreurs de saisie. La progression de la campagne peut être suivie par pays et par espèce sur une carte géographique. Chaque carte interactive présente les cellules où l'échantillonnage a commencé, celles où il est terminé, les positions des fermes et le nombre d'animaux dans chaque cellule. Les animaux peuvent également être sélectionnés en fonction de leurs caractéristiques (pays, espèce, race, cellule…). La structure de la base de données est présentée en détail à l'annexe A.

Chaque utilisateur dispose d'un compte personnel pour accéder au site Web. Les droits d'administration sont gérés par type d'utilisateurs, chaque groupe hérite des droits des groupes précédents : un échantillonneur (*sampler*) peut suivre l'évolution de la campagne — ces observateurs sont les partenaires dans les pays sans échantillonnage ; un rédacteur (*editor*) peut saisir dans la base de données des animaux issus de son pays et les modifier ; un chef (*head*) peut modifier tous les animaux dans son pays, désigner un chef d'équipe par cellule et marquer les cellules où l'échantillonnage est terminé ; finalement, un administateur (*administrator*) peut modifier toutes les entrées et créer de nouveaux comptes. L'interface est illustrée par la figure A.2.

4.2 Caractérisation géo-environnementale des habitats

L'environnement est décrit par deux jeux de variables climatiques et un jeu de données topographiques. Ces données sont mises à la disposition du public et directement utilisables pour la recherche.

4.2.1 Données de la « *Climate Research Unit* »

Le premier jeu de données est composé des moyennes mensuelles de huit variables environnementales (table 4.1). Elles sont produites et distribuées par la *Climate Research Unit* à Norwich (Royaume-Uni, http://www.cru.uea.ac.uk/) et couvrent tous les continents à l'exception de l'Antarctique (New et al., 2002). Les données proviennent des relevés de stations météorologiques terrestres entre 1961 et 1990. Le nombre de stations varie entre 3'952 pour la vitesse du vent et 27'075 pour les précipitations moyennes. Les valeurs sont interpolées sur une grille rectangulaire. La résolution spatiale est de 10' en longitude et en latitude, soit environ 18,5 km en Ouganda et 15,7 km au Maroc .

Variable	Description
wnd	moyennes mensuelles de la vitesse du vent en m/s, 10 mètres au-dessus du sol
dtr	moyennes mensuelles de l'amplitude thermique diurne en °C
frs	nombre de jours de gel par mois
pre	précipitations mensuelles en mm/mois
pre_sigma	valeurs du coefficient de variation des précipitations mensuelles
tmp	moyennes mensuelles de la température en °C
reh	moyennes mensuelles de l'humidité relative en pour cent
sunp	moyennes mensuelles du taux d'ensoleillement (en pour cent de la durée du jour)

Table 4.1 – Description des variables de la *Climate Research Unit*. Les valeurs sont disponibles pour chaque mois et sont complétées par une moyenne annuelle.

4.2.2 Données *WorldClim*

Le second jeu de données climatiques, *WorldClim - Global Climate Data*, couvre également tous les continents sauf l'Antarctique (http://www.worldclim.org/). Les relevés des stations météorologiques ont été restreints à la période 1950-2000 (si possible) avant d'être interpolés (Hijmans et al., 2005). Le nombre de stations varie entre 14'835 (températures minimum et maximum) et 47'554 stations (précipitations). Les données comportent quatre variables mensuelles et dix-huit variables dérivées (table 4.2). La résolution de la grille est de 30" en longitude et en latitude, soit environ 930 m en Ouganda et 790 m au Maroc. Les auteurs indiquent que les valeurs obtenues correspondent généralement bien à celles de la *CRU*, sauf au Groenland.

Variable	Description
tmin	Température minimale
tmean	Température moyenne
tmax	Température maximale
prec	Précipitations
BIO1	Température moyenne annuelle
BIO2	Amplitude diurne moyenne (Moyenne des valeurs mensuelles de (temp. max - temp. min))
BIO3	Isothermalité (BIO2/BIO7) ($\cdot$ 100)
BIO4	Saisonnalité de la température (écart type $\cdot$ 100)
BIO5	Température maximale du mois le plus chaud
BIO6	Température minimale du mois le plus froid
BIO7	Amplitude annuelle de la température (BIO5-BIO6)
BIO8	Température moyenne du trimestre le plus humide
BIO9	Température moyenne du trimestre le plus sec
BIO10	Température moyenne du trimestre le plus chaud
BIO11	Température moyenne du trimestre le plus froid
BIO12	Précipitations annuelles
BIO13	Précipitations du mois le plus humide
BIO14	Précipitations du mois le plus sec
BIO15	Saisonnalité des précipitations (Coefficient de variation)
BIO16	Précipitations du trimestre le plus humide
BIO17	Précipitations du trimestre le plus sec
BIO18	Précipitations du trimestre le plus chaud
BIO19	Précipitations du trimestre le plus froid

Table 4.2 – Description du jeu de données *WorldClim*. Les quatre premières variables sont disponibles par mois, alors que les dix-huit variables dérivées ont une valeur unique. Les températures sont indiquées en $[^{\circ}\mathrm{C} \cdot 10]$ et les précipitations en [mm].

4.2.3 Données *Shuttle Radar Topography Mission (SRTM)*

Les informations topographiques sont dérivées du modèle numérique d'altitude fourni par le *Shuttle Radar Topography Mission* (SRTM). Ce projet, mené par les agences spatiales américaine, allemande et italienne en 2000, a mesuré l'altitude du sol entre 60°de latitude nord et 56°de latitude sud avec deux radars embarqués à bord de la navette *Endeavour*.

Le modèle numérique mondial que j'ai utilisé a une résolution de 3', soit environ 93 m en Ouganda et 79 m au Maroc.

4.3 Extraction de l'information génétique

Cette section donne un aperçu des deux méthodes utilisées par NextGen pour déchiffrer le code génétique ainsi que du filtrage des données précédant leur analyse. Les échantillons collectés sur le terrain sont plongés dans l'alcool durant une journée puis placés dans un gel de silice, ce qui permet de les conserver à température ambiante.

4.3.1 Séquençage intégral

Le séquençage concerne les 388 chèvres et moutons échantillonnés au Maroc et en Iran ainsi que 25 bovins d'Ouganda. Ces analyses sont effectuées par le Génoscope (Centre National de Séquençage, Paris), partenaire NextGen, au moyen du séquenceur « HiSeq 2000 » (Illumina Inc., San Diego, CA). Pour commencer, l'ADN est extrait des cellules, purifié, puis coupé en fragments d'environ 800 paires de bases (*base pairs*, bp). Ces fragments doivent être amplifiés (multipliés) afin d'obtenir un signal lumineux suffisant lors de la lecture (cf p. 17).

Amplification par accrochage-liaison sur puce (*bridge amplification*)

Des adaptateurs (courtes séquences d'ADN prévues pour les connexions) sont fixées à chaque extrémité des fragments et ceux-ci sont dénaturés (le double brin d'ADN est séparé en deux brins simples). L'analyse se déroule sur une puce (*flow-cell*) à laquelle sont fixés de nombreux adaptateurs. Les fragments (brins simples) mesurant entre 150 et 200 bp sont déposés sur la puce et s'y accrochent par leurs extrémités. Comme ils possèdent deux adaptateurs, ils se courbent et forment de petits « ponts ». Puis une solution de polymérases et de nucléotides est versée sur la puce et les enzymes synthétisent les brins complémentaires d'ADN. La solution est ensuite lavée et les fragments sont dénaturés. Une de leurs extrémités est détachée de la puce et se recombine à un autre adaptateur. L'amplification se poursuit jusqu'à former 100 à 200 millions de grappes de brins d'ADN (*template clusters*) fixés par une extrémité à la surface de la puce.

Séquençage par synthèse (*sequencing by synthesis*)

Une amorce est fixée à l'extrémité de chaque brin. Lors de la synthèse, chaque nucléotide est lié à un terminateur de chaîne réversible marqué par un fluorochrome. Ainsi les polymérases ne peuvent synthétiser qu'un nucléotide par brin. Un laser excite les terminaisons fluorescentes qui sont identifiées par la couleur du rayonnement émis. Les terminaisons sont ensuite décrochées et une nouvelle base est fixée sur chaque brin. La synthèse se poursuit jusqu'à lire 100 paires de bases. Comme les fragments d'ADN simples brins fixés à la puce mesuraient environ 180 pb, les séquences lues sur les brins complémentaires obtenus après l'amplification se chevauchent et peuvent être réunies, voir fig. 4.2.

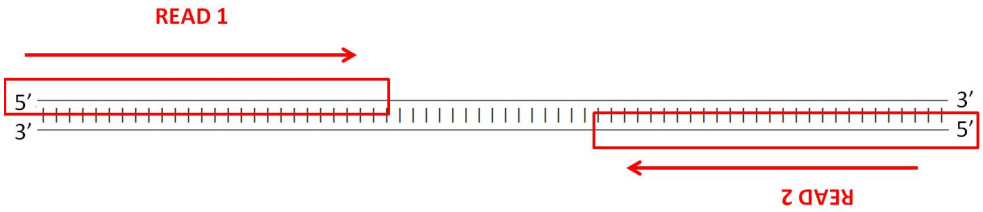


Figure 4.2 – Séquençage par extrémités appairées [1].

Si les fragments d'ADN analysés sont plus longs, le séquenceur lira 100 pb à chaque extrémité et les *reads* auront un intervalle inconnu au milieu. Il est très courant de séquencer plusieurs librairies avec des fragments de différentes tailles. Les *reads* « troués » fournissent des informations sur des régions éloignées les unes des autres sur un chromosome et facilitent l'assemblage du génome.

Assemblage et alignement

Le séquençage Sanger produit des *reads* suffisamment longs pour que l'assemblage se base sur un test d'alignement de toutes les paires de *reads* possibles. Les séquenceurs de seconde génération produisent de très nombreux *reads* très courts et cette méthode n'est plus applicable. D'autres techniques ont été développées reconstituer des génomes longs, comme ceux des mammifères, à partir de *reads* courts.

Butler et al. (2008) proposent `ALLPATHS`, un algorithme pour l'assemblage de génomes courts. Cette approche utilise des *reads* courts appairés issus du séquençage de deux bibliothèques, les fragments mesurent environ 200 bp et 4k bp (MacCallum et al., 2009). Un « K-mer » est une séquence de nucléotides de longueur K existant sur au moins un *read*. `ALLPATHS` crée un répertoire de tous les K-mer d'une taille donnée (par ex. $K = 20$). Deux k-mers sont adjacents si les $K-1$ derniers nucléotides du premier sont identiques aux $K-1$ premiers nucléotides du second. L'algorithme aligne les K-mers adjacents en « *uni-paths* », les plus longues séquences univoques possibles. `ALLPATHS` crée ensuite un graphe avec les « *uni-paths* ». Les noeuds du graphe sont les loci ambigus où plusieurs « *uni-paths* » sont compatibles. La bibliothèque à 4k

1. Illustration : Minikel, 2012

bp fournit des informations sur des séquences éloignées, ce qui facilite la construction des « *uni-paths* ». `ALLPATHS-LG` étend cette approche à l'assemblage de génomes longs (Gnerre et al., 2011). Les améliorations concernent la vérification approfondie des erreurs de séquençage, la gestion des séquences répétées ainsi que des régions où le taux de couverture est faible. Le séquençage du premier individu d'une nouvelle espèce est généralement effectué avec un taux de couverture élevé pour faciliter l'assemblage et réduire le risque d'erreur. Etant donné que la plus grande partie du génome ne varie pas entre individus de la même espèce, les échantillons suivants sont plus faciles à assembler, car ils peuvent être alignés sur ce « génome de référence ».

4.3.2 Génotypage

Le séquençage et la comparaison du génome de plusieurs individus de la même espèce permet d'identifier les loci où le génotype peut varier. Les polymorphismes nucléotidiques (*single nucleotide polymorphisms*, SNPs) sont des loci variables dont l'allèle mineur (l'allèle ayant la fréquence la plus petite) a une fréquence suffisante pour perdurer au sein de la population. Un locus peut être localisé sur le génome par la séquence de nucléotides située directement avant ou après. Si la séquence est suffisamment longue (par ex. 50 bp), le locus sera identifié sans ambiguité.

Les puces à ADN pour le génotypage comportent de nombreux brins d'ADN à leur surface et chaque brin s'hybride à une séquence voisine d'un SNP. Le dernier nucléotide du brin est le complément d'un des allèles possibles du SNP (Gunderson et al., 2005). Plusieurs brins du même type sont placés sur des billes à la surface de la puce pour augmenter le signal. Les puces de génotypage détectent des SNPs bialléliques : deux types de sondes déchiffrent les deux nucléotides possibles pour chaque locus ciblé. Les fragments à analyser sont préparés en leur fixant un marqueur fluorescent et placés sur la puce. Après la réaction, la puce est scannée avec un laser pour mesurer l'intensité lumineuse de chaque point. Chaque locus homozygote présente un seul point lumineux car toutes les séquences se sont fixées au même type de brin. La position du point permet de connaître l'allèle. Les loci hétérogyzotes sont identifiés par deux points lumineux, car les séquences se sont réparties entre les deux types de brins en fonction de leur allèle.

4.3.3 Filtrage

En désignant par A et G les nucléotides lus pour un SNP, les trois génotypes possibles à chaque locus sont AA, AG et GG. Les SNPs ne sont pas phasés : la combinaison GA n'est pas distinguable de AG. Les données brutes contiennent une ligne par individu et deux colonnes par locus. Lorsque qu'un locus n'a pas été déchiffré pour un échantillon, les deux nucléotides correspondants sont indiqués comme manquants.

Les données ont été filtrées en trois étapes avec le programme `PLINK` en vue de leur analyse

(Purcell et al., 2007 ; Purcell, 2009).

1. Le taux de génotypage - le ratio entre le nombre de loci lus et le nombre total - est calculé pour chaque individu (*per-individual genotyping rate*). Les échantillons dont le taux est inférieur au seuil fixé sont écartés de l'analyse.
2. Le taux de génotypage par SNP (*SNP missing genotype rate*) est calculé pour les échantillons restants. Les loci avec un taux inférieur au seuil sont également exclus.
3. La fréquence de l'allèle mineur A ou G (*minor allele frequency*, MAF) détermine quels loci sont considérés comme polymorphiques dans la population. Les allèles très rares ne sont généralement pas utilisés pour l'étude de la diversité génétique, car ils apportent peu d'information. Le taux est souvent fixé à 5%, ce qui correspond à 0,25% d'individus homozygotes et 9,5% d'hétérozygotes si la population est en équilibre de Hardy-Weinberg.

Les trois seuils de sélection peuvent être fixés indépendamment.

4.4 Choix des échantillons au Maroc

La campagne d'échantillonnage a permis de réunir des tissus cutanés de 1'283 chèvres [2] et 1'412 moutons (voir fig. 4.3). Cependant les ressources financières n'ont permis de séquencer que 164 individus pour chaque espèce. Il a donc été nécessaire de choisir les individus de manière à optimiser l'information tirée des analyses génétiques (en relation ou non avec l'environnement) tout en tenant compte de trois contraintes.

— La variabilité climatique des habitats correspondants doit être maximisée pour capturer la variabilité génétique potentiellement adaptative des animaux.

— Les échantillons doivent être répartis sur tout le territoire (représentativité spatiale).

— Chaque race doit avoir suffisamment de représentants afin d'assurer la significativité des statistiques lors des analyses.

La sélection des individus a été menée indépendamment sur les chèvres et les moutons [3]. Notre approche a consisté à choisir 164 fermes pour assurer une bonne représentativité spatiale et environnementale puis à tirer aléatoirement un animal par ferme et à vérifier que la sélection contient suffisamment d'individus de chaque race. Elle se déroule en trois étapes : une analyse en composantes principales, une classification ascendante hiérarchique et un tirage aléatoire d'un individu par classe.

4.4.1 Méthode de sélection

Les échantillons à séquencer doivent être représentatifs de la variabilité génétique et écologique des petits ruminants marocains. Les individus doivent donc être répartis de manière

2. Dix-sept chèvres étaient issues d'un croisement avec une race espagnole et ont été écartées de l'analyse.

3. Cette analyse a été réalisée avec la collaboration avec Diane Perez dans le cadre du cours de plans d'expériences (*Design of Experiments*) de Jean-Marie Fürbringer.

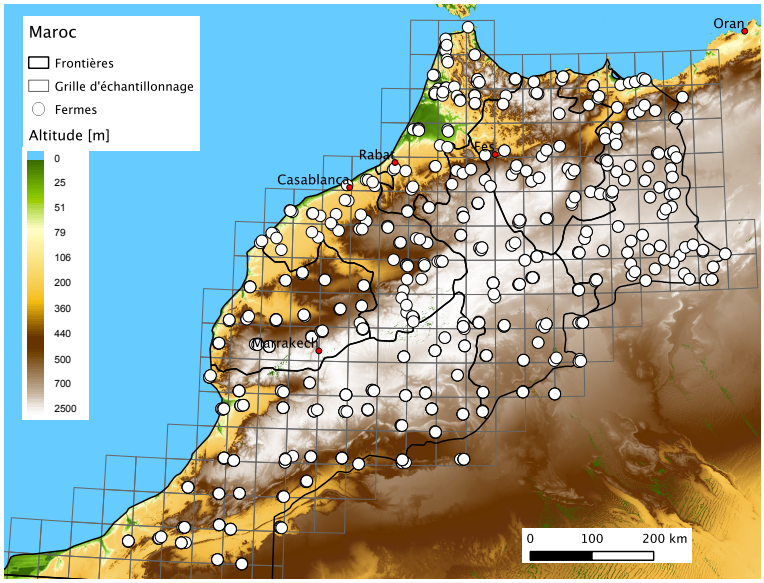


Figure 4.3 – Cartes des 577 fermes élevant des chèvres (412) et/ou des moutons (432) au Maroc. Pour chaque espèce, trois fermes ont été visitées dans chaque cellule de la grille d'échantillonnage et trois animaux ont été échantillonnés dans chaque ferme.

continue le long des gradients environnementaux pour échantillonner régulièrement l'espace des paramètres climatiques. En effet, la prise en compte des extrema climatiques ne fournit pas l'information adéquate pour calibrer des modèles linéaires généralisés et n'a pas d'utilité pour la génétique des populations. Comme les animaux présents dans une ferme partagent le même habitat, nous avons utilisé les fermes comme unités élémentaires lors de la classification.

Analyse en composantes principales

Une analyse en composantes principales (ACP) consiste à transformer K variables corrélées (les variables environnementales) en K facteurs non-corrélés, les *composantes principales* ou *axes principaux* (Escoffier et Pagès, 2008). Ces composantes sont obtenues par rotation des variables originales, en diagonalisant la matrice des covariances. Le premier axe représente la direction le long de laquelle les points varient le plus. Les axes suivants expliquent tour à tour le maximum de variance restante, tout en étant orthogonaux entre eux. Les composantes sont donc ordonnées en fonction de la part de variance qu'elles expliquent.

L'ACP des variables environnementales fournit une image synthétique du climat. Elle nous permet de choisir des échantillons « aussi différents que possible » : les premiers axes principaux expliquent l'essentiel de la variabilité climatique tout en limitant le bruit aléatoire. La distance euclidienne dans ce sous-espace permet de définir une distance écologique entre les fermes. Cela limite le poids accordé aux facteurs très corrélés lors de la sélection des échantillons à

séquencer. Comme les variables environnementales sont de nature et d'amplitude numérique différentes, elles sont centrées et réduites avant de calculer l'ACP :

$$\tilde{X}_k = \frac{X_k - \bar{X}_k}{\sigma(X_k)} \tag{4.1}$$

où $\bar{X}_k$ est la moyenne et $\sigma(X_k)$ est l'écart-type de la distribution.

Classification ascendante hiérarchique

Le choix d'un nombre limité de points représentant la diversité des habitats consiste à regrouper les fermes en fonction de leurs similitudes (basées sur les distances écologiques) et d'en choisir une par groupe. Nous avons utilisé une classification ascendante hiérarchique avec le critère de Ward. C'est une méthode itérative où les deux éléments les plus proches (points et/ou groupes) sont réunis à chaque étape. L'algorithme de Ward choisit quels élements fusionner en minimisant l'augmentation de l'inertie intra-classe. Ce critère tend à réunir les classes proches contenant peu d'éléments (Escoffier et Pagès, 2008, p. 46). La classification ascendante hiérarchique est déterministe et les partitions générées à chaque itération sont emboîtées. C'est pourquoi le processus est souvent représenté sous la forme d'un arbre.

Tirage sous contraintes

La stratégie d'échantillonnage basée sur une grille fournit une densité régulière de points sur le territoire. Cependant le choix aléatoire d'une ferme par classe est susceptible d'introduire un biais spatial. C'est pourquoi la représentativité spatiale des échantillons sélectionnés est quantifiée avec un indice de répartition (D). Nous l'avons défini comme la somme des distances entre chaque point et son plus proche voisin dans la sélection. Plus l'indice est élevé, plus les points sont éloignés les uns des autres et meilleure est leur répartition spatiale.

4.4.2 Choix des échantillons

Nous avons utilisé les données CRU pour cette analyse. La résolution des cellules de la grille (10') permet d'avoir un seul fichier par variable environnementale, ce qui simplifie leur traitement. J'ai transformé ces données discrètes en surfaces avec la méthode de Voronoi fournie par le logiciel SIG `Manifold`[4]. Comme les 117 couches environnementales sont difficiles à charger d'un bloc dans ce logiciel, j'ai extrait au préalable les points compris dans deux zones recouvrant le Maroc et l'Ouganda avec un script ad-hoc. J'ai ensuite assigné à chaque ferme les valeurs climatiques de la cellule de Voronoi correspondante.

La suite des analyses a été effectuée dans `R` (R Core Team, 2013). L'ACP calculée sur les $K = 117$ variables environnementales a mis en évidence leur forte corrélation : les sept premiers axes

4. www.manifold.net

représentent 96% de la variance. La table 4.3 montre les coordonnées des variables originales le long des deux premiers axes.

Composante principale 1		Composante principale 2	
Nom	Projection	Nom	Projection
pre_sigma.1	-0.12588	dtr.7	0.13723
pre_sigma.2	-0.12941	dtr.8	0.13855
pre_sigma.3	-0.12600	frs.1	0.15954
pre.2	0.12625	frs.11	0.14393
pre.3	0.12429	frs.12	0.15803
rd0.1	0.12977	frs.2	0.15402
rd0.10	0.12357	frs.3	0.14827
rd0.11	0.12856	frs.4	0.13886
rd0.12	0.12956	frs.avg	0.15593
rd0.2	0.13118	pre.7	0.14041
rd0.3	0.13201	pre.8	0.15722
rd0.4	0.12867	pre.9	0.15022
rd0.avg	0.12656	rd0.7	0.15294
reh.4	0.12359	rd0.8	0.16157
sunp.1	-0.12485	rd0.9	0.16062
sunp.11	-0.13004	tmp.1	-0.16573
sunp.12	-0.12688	tmp.11	-0.15915
sunp.2	-0.12813	tmp.12	-0.16638
sunp.3	-0.12829	tmp.2	-0.15598
sunp.4	-0.12875	tmp.3	-0.13678

Table 4.3 – Projections des variables environnementales sur les deux premiers axes principaux. Les 20 plus fortes corrélations sont regroupées par famille de variables.

Ces axes expliquent respectivement 44% et 30% de la variance totale.

Les axes principaux ne sont pas liés à une variable environnementale en particulier, on peut cependant relever certaines tendances. Le premier axe est corrélé positivement avec le nombre de jours pluvieux et négativement avec l'ensoleillement et la variation des précipitations. La fig. 4.4 montre que cet axe différencie les régions côtières des désertiques au Maroc. Le second axe principal est associé avec une grande amplitude de températures diurnes, le nombre de jours de gel, des précipitations abondantes et de basses températures hivernales. Cette composante est corrélée avec l'altitude, comme illustré sur la fig. 4.5.

Nous avons retenu les sept premiers axes principaux, qui expliquent un total de 96% de la variance et au minimum 1% de la variance individuellement. L'espace généré par ces axes représente de manière fiable les conditions climatiques. La distance euclidienne dans cet espace permet de définir une distance environnementales entre les fermes en vue de leur agrégation.

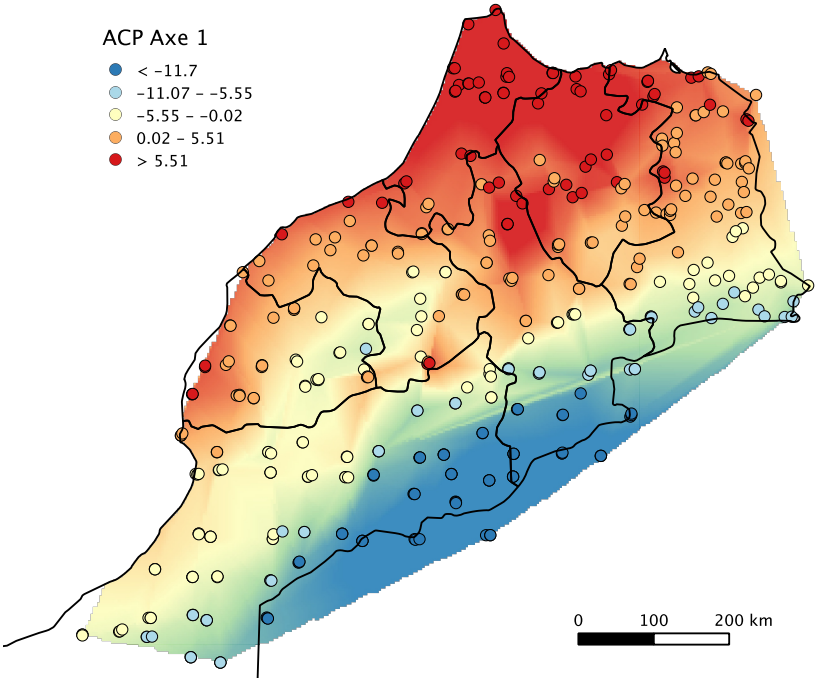


Figure 4.4 – La première composante principale distingue les régions côtières des régions désertiques du Maroc. Les valeurs élevées (en rouge) sont corrélées avec le nombre de jours de pluie tandis que les valeurs basses (en bleu) se situent près du désert où l'ensoleillement est important et où les précipitations varient peu. L'interpolation des valeurs est présentée en arrière-plan pour faciliter la visualisation de leur distribution spatiale. Modèle numérique d'altitude : SRTM3.

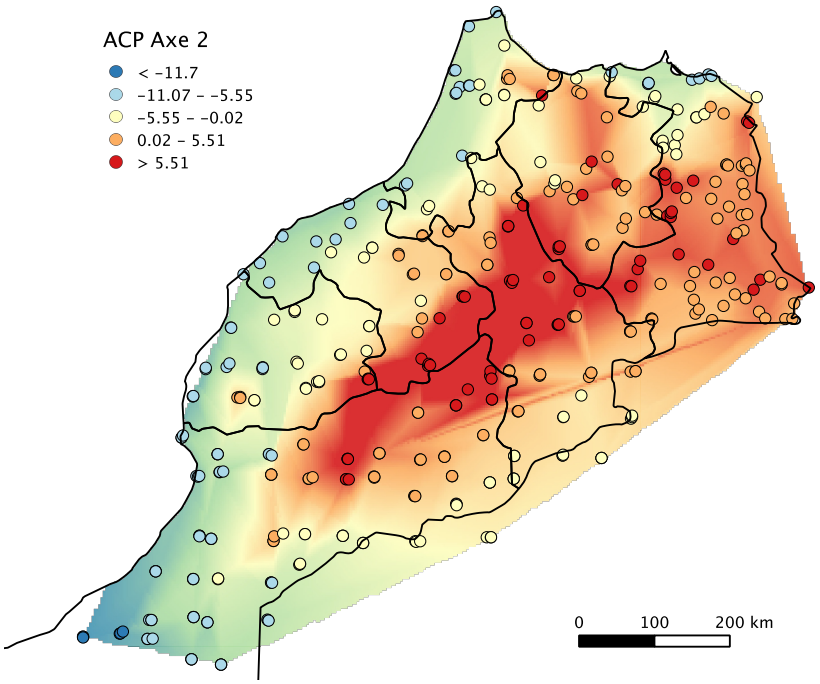


Figure 4.5 – La deuxième composante principale est corrélée avec l'altitude. Les valeurs élevées (en rouge) se situent dans la chaîne de l'Atlas où la température hivernale est basse, alors que l'amplitude de température diurne, le nombre de jours de gel et les précipitations y sont élevés. L'interpolation des valeurs est présentée en arrière-plan pour faciliter la visualisation de leur distribution spatiale. Modèle numérique d'altitude : SRTM3.

Parmi les fermes visitées, 412 comprennent des chèvres et 432 des moutons. Certaines fermes se situent dans la même cellule climatique et ont été groupées pour l'analyse, ce qui mène à 226 conditions climatiques différentes pour les chèvres et 229 pour les moutons. Ces groupes de fermes représentent des points distincts dans l'espace des conditions climatiques et ont servi d'unités élémentaires pour la classification ascendante hiérarchique. La fig. 4.6 présente la répartition des fermes en 10 classes déterminées avec la CAH.

La sélection des individus utilise le niveau d'agrégation des fermes en 164 classes. La figure 4.7 montre la distribution des points climatiques pour quatre variables environnementales. Comme attendu, chaque classe présente une faible amplitude de valeurs climatiques.

Chaque classe contient entre 1 et 3 fermes, soit entre 1 et 18 individus. Pour optimiser la représentativité spatiale, nous avons effectué 50 tirages aléatoires d'un individu par classe et avons conservé celui qui présentait l'indice de répartition spatiale (D) le plus élevé. Le processus a été répété 10 fois et nous avons plus constater que l'indice de répartition D (qui est le maximum de 50 valeurs) et le nombre d'individus par race variaient peu entre les itérations. Le tirage final a été choisi parmi ceux présentant une haute valeur de l'indice D et le maximum d'animaux n'appartenant à aucune race. L'échantillon retenu contient suffisamment d'animaux de chaque race, tout en conservant suffisamment d'animaux « sans race ». Cette précaution est prise pour représenter la diversité génétique des petits ruminants marocains au cas où certaines races auraient suivi des programmes de sélection.

4.4.3 Sous-échantillonnage

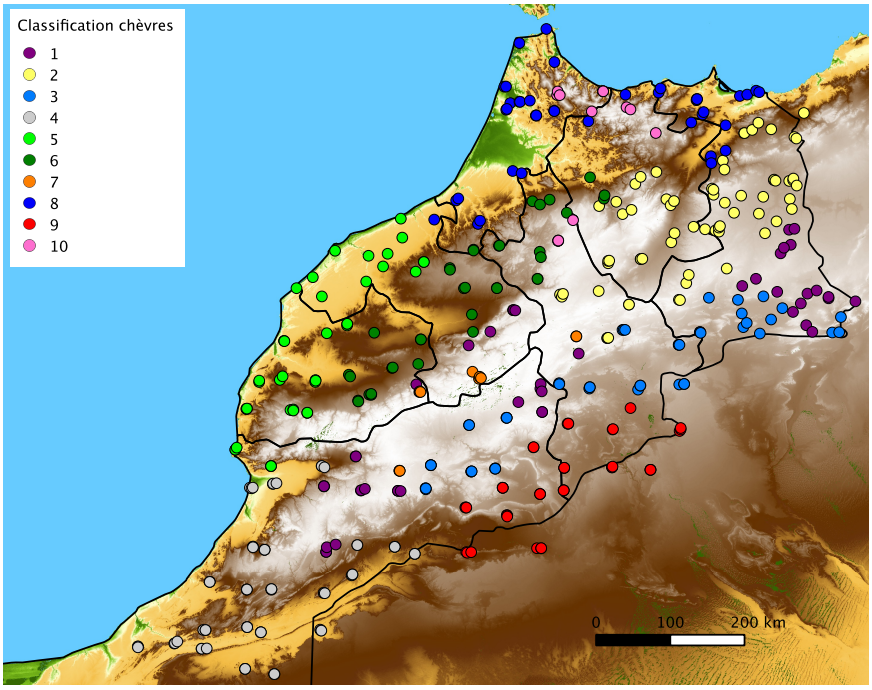
Cependant des contraintes supplémentaires sont apparues après le choix des individus (voir fig. 4.8). Le taux de couverture retenu pour le séquençage était susceptible d'impliquer une nouvelle analyse pour certains échantillons. Les animaux devaient donc être répartis en deux groupes afin de pouvoir reséquencer le premier et ainsi doubler le taux de couverture si nécessaire. Le premier lot de 82 échantillons devait ainsi former un sous-ensemble représentatif de la population.

De plus, trente individus de chaque espèce devaient être génotypés avec des puces 50k afin de comparer les résultats avec ceux du séquençage. Il était nécessaire de tirer ce sous-échantillon dans le premier groupe sélectionné, tout en s'assurant d'avoir au moins trois individus par race. Certains échantillons ne contenant pas assez de matériel génétique pour les deux types d'analyses devaient être assignés au deuxième groupe de séquençage.

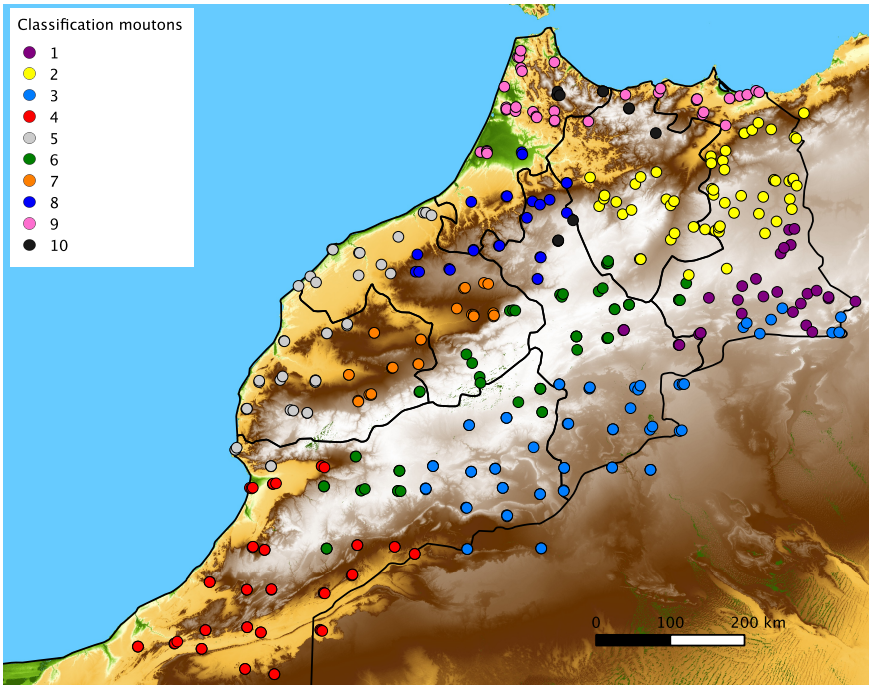
Par définition, la classification ascendante hiérarchique produit des résultats emboîtés. En stoppant le processus de regroupement à 82 classes, les 164 classes précédemment définies forment des groupes de un à quatre éléments. La situation est similaire à la précédente, 164 individus sont répartis en 82 classes. La méthode de tirage sous contrainte est appliquée pour choisir un échantillon par classe en maximisant la distribution spatiale et la représentativité des races parmi 50 répétitions de 50 tirages.

Pour le choix des échantillons à génotyper, il n'était pas possible de tenir compte à la fois de la classification en 30 groupes, de la répartition spatiale et de la distribution des races. Les tirages sont effectués en deux phases parmi les échantillons du premier lot de séquençage : Trois animaux sont choisis aléatoirement pour chaque race, puis les autres échantillons sont tirés sans tenir compte des races. La distribution spatiale est maximisée sur 50 tirages et le processus est répété 50 fois. Le choix final s'est arrêté sur un tirage avec un indice spatial élevé et un nombre d'individus par race représentatif de la population échantillonnée.

Les positions géographiques des fermes pour les trois sélections imbriquées sont illustrées sur la fig. 4.9. La fig. 4.10 montre la répartition des fermes le long des sept axes principaux de l'ACP. La méthode présentée ici permet de choisir quels échantillons analyser en fonction des caractéristiques environnementales des habitats lorsque les ressources sont limitées. La classification sert à lisser la distribution environnementale des échantillons. Elle s'apparente à une sélection par stratification utilisée pour sélectionner des échantillons en s'affranchissant des biais lors de la récolte.



(a) Fermes élevant des chèvres



(b) Fermes élevant des moutons

Figure 4.6 – Positions des fermes élevant des petits ruminants. Les couleurs représentent la classification en fonction des conditions climatiques. Comme escompté, les fermes sont globalement regroupées en fonction de leur proximité spatiale et de l'altitude. La classification distingue également les régions côtières de l'intérieur des terres. Modèle numérique d'altitude : SRTM3.

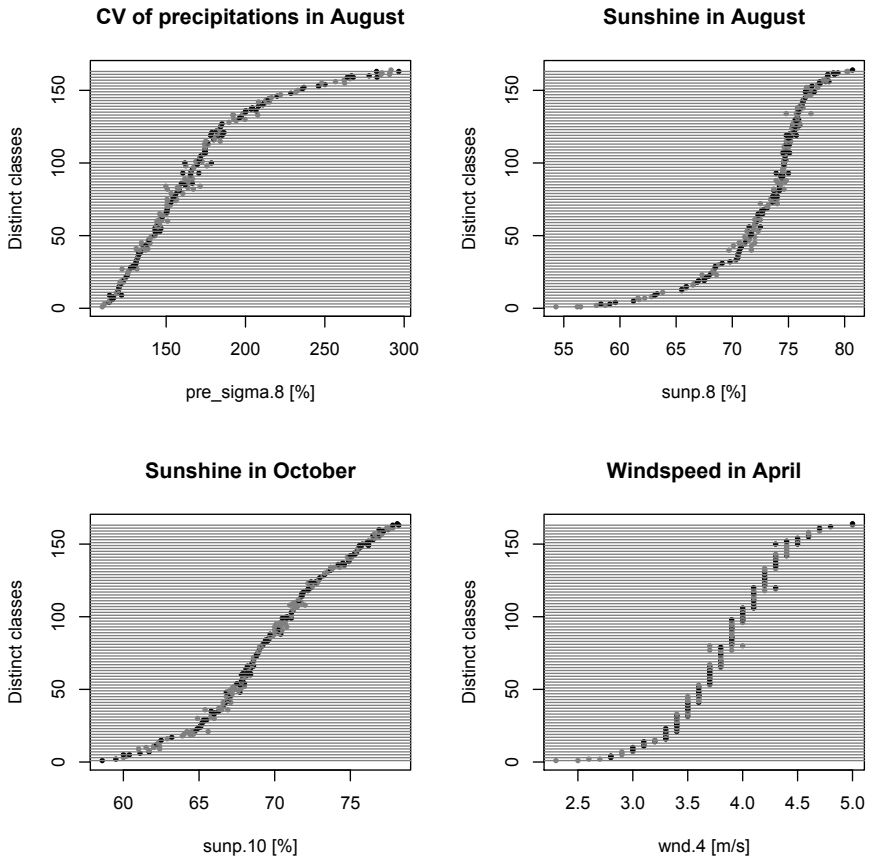


Figure 4.7 – Les 164 classes sont représentées en fonction de quatre variables environnementales d'origine. Chaque ligne montre les valeurs prises par les fermes appartenant à une classe. Par souci de lisibilité, les classes ont été triées en fonction de la valeur moyenne des fermes.

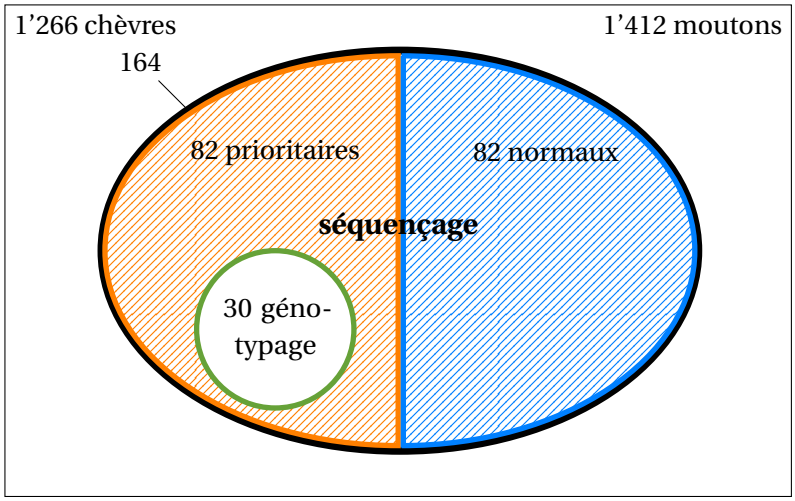
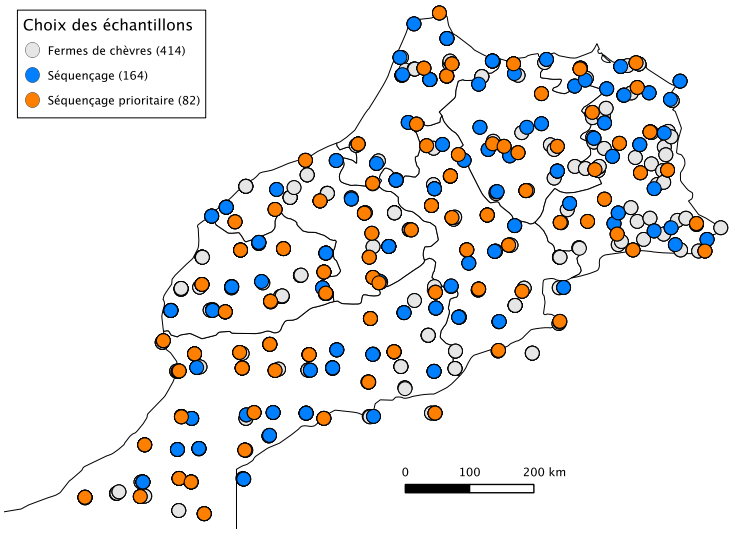
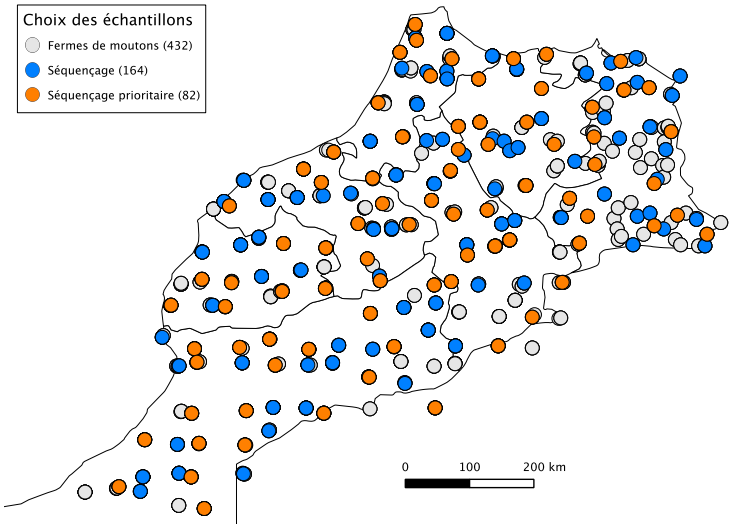


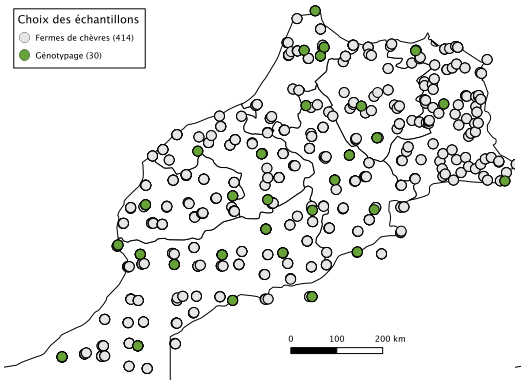
Figure 4.8 – Le protocole du projet prévoyait de séquencer 164 animaux de chaque espèce au Maroc (ellipse noire). Cependant, la qualité des données obtenues aurait pu nécessiter de réanalyser certains échantillons (en cas de taux de couverture insuffisant). C'est pourquoi les échantillons de chaque espèce ont été séparés en deux groupes. Le premier groupe de 82 individus serait analysé en priorité (figure orange). Si la qualité des données obtenues était satisfaisante, le deuxième groupe serait séquencé à son tour (figure blue). De plus, 30 échantillons seraient également génotypés à des fins de comparaisons. Ces individus devaient donc être choisis parmi le premier groupe de séquençage (figure verte).



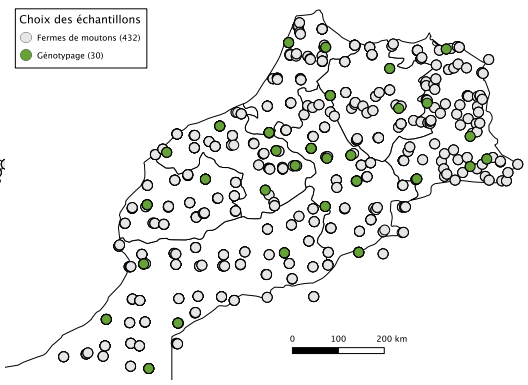
(a) Séquençage des chèvres



(b) Séquençage des moutons



(c) Génotypage des chèvres



(d) Génotypage des moutons

Figure 4.9 – *a*) et *b*) Cartes des fermes où les animaux ont été sélectionnés pour le séquençage intégral. La sélection du sous-ensemble de 82 individus (points oranges) est incluse dans l'ensemble des 164 individus (points oranges et bleus). *c*) et *d*) Les sous-ensembles de 30 échantillons ont été sélectionnés dans les groupes de 82 individus prioritaires pour du génotypage par puce ADN de 50k SNPs.

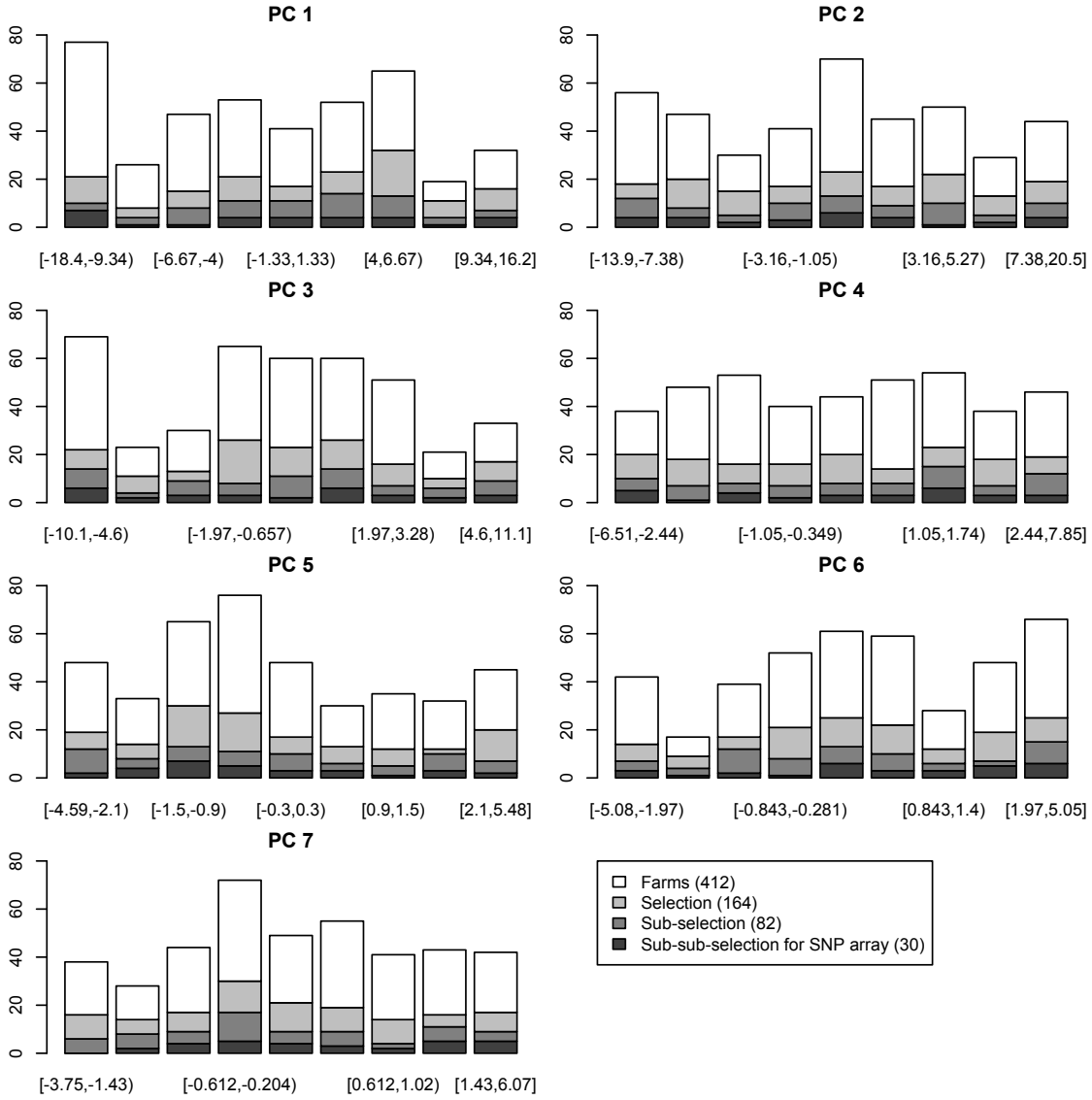


Figure 4.10 – Distribution des fermes de chèvres sur les axes principaux. Les trois ensembles d'individus sélectionnées sont imbriqués.

4.5 Données disponibles en Ouganda

La situation en Ouganda diffère de celle du Maroc. D'une part la quasi-totalité des échantillons (915 sur 917) ont été génotypés avec une puce comprenant 54k ou 800k SNPs. Des données moléculaires sont donc disponibles pour l'ensemble du territoire. Les conditions climatiques y sont en outre plus homogènes qu'au Maroc. La sélection des 102 individus pour la puce à haute densité repose essentiellement sur leur représentativité spatiale, ce qui est moins contraignant que dans le cas précédent. Le partenaire NextGen responsable des analyses pour l'Ouganda a donc choisi deux individus dans chacune des 51 cellules.

La distribution spatiale des animaux ougandais est illustrée sur la figure 4.11 et les différentes races échantillonnées sont présentées à la table 4.4

Race	Animaux
ankole	339
ankole Zebu cross	11
East African Shorthorn Zebu	153
Nganda	41
Nkiga	32
Nsongora	11
Ntoro	10
Ntuku	2
Shorthorn Zebu	243
Small East African Zebu	75
Total	917

Table 4.4 – Décompte des bovins échantillonnés en Ouganda. Les données récoltées comprennent la race et quelques caractéristiques morphologiques de chaque individu.

4.5.1 Données moléculaires

Les animaux échantillonnés en Ouganda ont été séparés en deux groupes : 813 bovins ont été génotypés avec la puce *BovineSNP50 BeadChip* qui comprend 54'609 SNPs et 102 autres animaux avec la puce *BovineHD BeadChip* de 777'692 SNPs (Illumina Inc., San Diego, CA). Ces jeux de données seront désignés par les abréviations « 54k » et « 800k ».

Les données moléculaires ont été filtrées avec le logiciel `PLINK` en utilisant un taux de réussite de 95% pour les SNPs et les individus [5]. Les deux principaux jeux de données utilisés sont les suivants :

Données 54k soit 804 individus et 41'215 SNPs, fréquence allélique minimale (M. A. F.) : 1% ;

Données 800k soit 102 individus et 634'849 SNPs, M. A. F. : 5%.

5. L'analyse des données ougandaises a été réalisée en collaboration avec Pablo Orozco-terWengel. Notre travail a été grandement facilité par mon séjour dans le groupe « Organismes et Evolution » du Prof. Mike Bruford à l'Université de Cardiff, groupe auquel Pablo appartient. (Bourse European Science Foundation n° 4118)

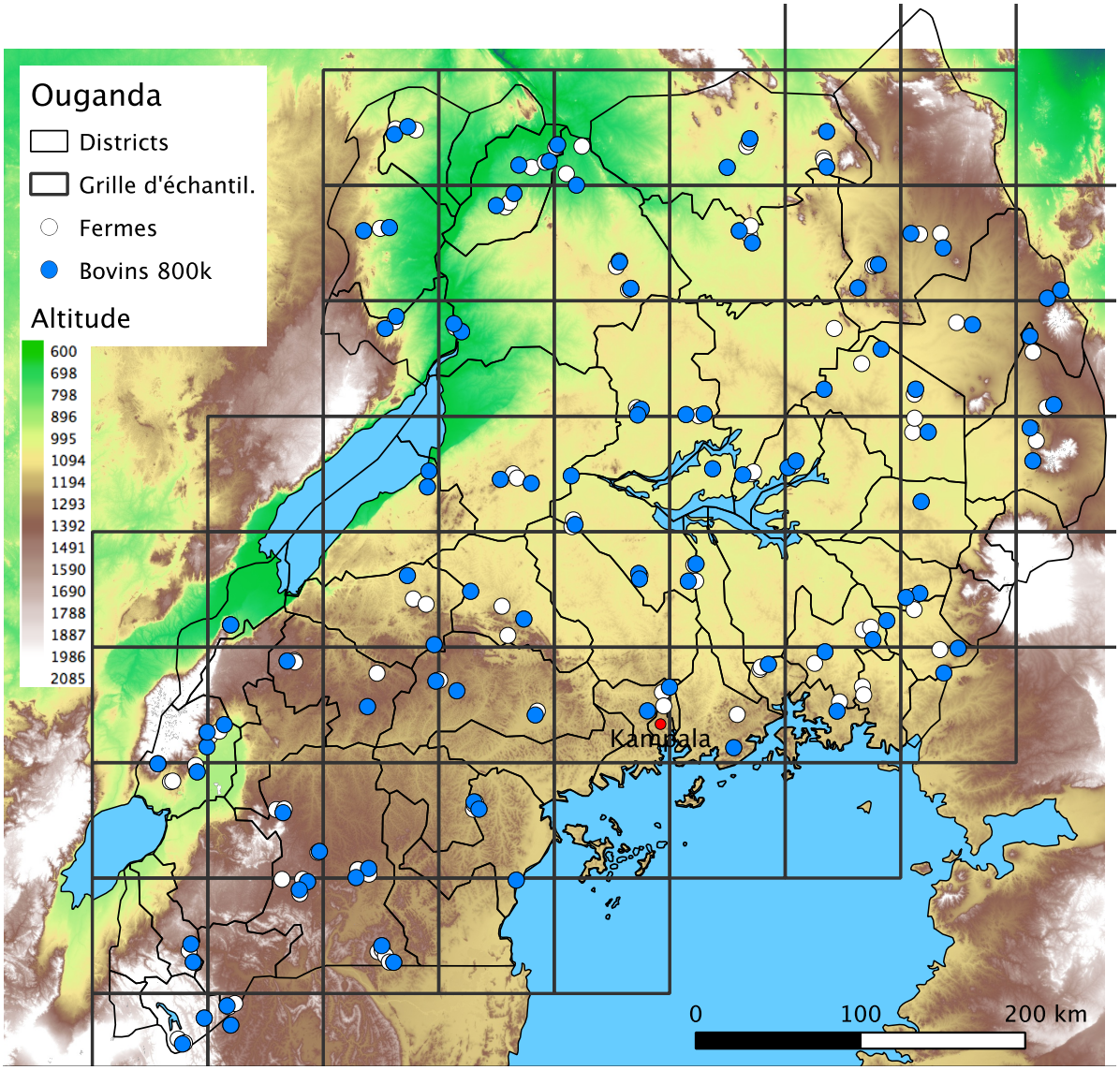


Figure 4.11 – Fermes où des bovins ont été échantillonnées en Ouganda. En général, quatre individus ont été choisis dans chaque ferme. Chaque cellule contient quatre fermes et le pas de la grille est d'environ 70 km. Les points bleus indiquent les fermes où un animal a été choisi pour le génotypage à haute densité.

4.5.2 Préparation des variables environnementales

La détection de signatures de la sélection naturelle en génomique environnementale nécessite de caractériser les habitats des organismes étudiés. La sélection des variables environnementales est une étape cruciale de l'analyse car ce sont ces variables qui représentent la pression de sélection dans les modèles corrélatifs. Les données présentées ici seront utilisées au chapitre 7 pour construire les modèles uni- et bivariés de Samβada, ainsi que ceux de `BayEnv` et `LFMM`. J'ai sélectionné les variables environnementales parmi les données climatiques WorldClim et parmi des variables topographiques dérivées de SRTM. Ces données de haute résolution sont disponibles sous forme de « tuiles » qu'il faut juxtaposer pour couvrir la région désirée. Les données WorldClim concernant l'Ouganda sont réparties sur quatre de ces tuiles. Les fichiers étant disséminés dans de nombreux sous-dossiers, l'importation et l'assemblage manuels des 67 variables dans un logiciel SIG auraient été fastidieux, c'est pourquoi j'ai écrit un script en Python qui colle les tuiles correspondant à chaque variable avec la bibliothèque `gdal` [6].

Pour coller les 36 tuiles du modèle numérique d'altitude SRTM et dériver la pente et l'orientation du terrain en Ouganda, j'ai utilisé le logiciel `SAGA` [7] qui comprend de nombreux modules d'analyse spatiale.

L'extraction des données environnementales pour chaque individu (« *spatial overlay* ») a quant à elle été réalisée avec le logiciel `QuantumGIS` (`QGIS`) [8]. Les informations utilisées pour l'analyse sont les précipitations, les températures mensuelles (minimales, moyennes et maximales), 19 variables bioclimatiques dérivées, la longitude, la latitude, l'altitude, la pente et l'orientation, soit 72 variables environnementales (voir table 4.2).

Les trois programmes sus-mentionnés sont libres (*open source*).

Ces 72 variables environnementales sont corrélées, premièrement en raison de leur résolution temporelle (deux mois consécutifs présentent souvent des températures et des précipitations similaires) et deuxièmement pour des raisons géographiques. En effet, l'Ouganda est situé sur l'équateur et les conditions climatiques y suivent un gradient nord-sud. Certaines variables environnementales sont ainsi très corrélées avec la latitude.

Ces corrélations entre variables environnementales créent plusieurs problèmes lors de la détection de signature de sélection. Premièrement, les tests statistiques utilisés sont prévus pour des modèles indépendants, ce qui n'est pas le cas si certaines variables explicatives sont corrélées. Deuxièmement, la corrélation entre variables prédictrices dans un modèle multivarié augmente la variance de l'estimation des paramètres de la régression. Ce phénomène appelé « multicollinéarité » est source d'instabilité lors de la calibration de modèles multivariés, l'ajout ou le retrait de quelques points échantillonnés pouvant alors modifier radicalement les paramètres estimés. Troisièmement, lorsque des variables environnementales sont très corré-

6. www.gdal.org
7. www.saga-gis.org
8. www.qgis.org

lées, les modèles construits à partir de ces variables génèrent des distributions de présence très proches pour un marqueur génétique, au point que certains de ces modèles peuvent être considérés comme équivalents. Au vu du contexte statistique et du fait que le temps de calcul des modèles corrélatifs est proportionnel au nombre de variables environnementales considérées, la sélection d'un sous-ensemble de variables permet, d'une part, de limiter les erreurs d'analyse liées aux corrélations et, d'autre part, d'accélérer l'analyse des données.

Afin de limiter les corrélations entre variables environnementales, j'ai préparé deux sous-ensembles à partir des 72 variables disponibles : un pour les analyses univariées avec Samβada, `BayEnv` et `LFMM` (voir paragraphe suivant) et l'autre pour les analyses bivariées avec Samβada (voir p. 54).

Pour les modèles corrélatifs univariés

Le choix des variables est basé sur les données environnementales qui caractérisent l'habitat des 804 individus retenus dans le jeu de données de 54k SNPs. Les données sont constituées de la valeur prise par chacune des 72 variables environnementales à la position où a été échantillonné chaque animal. Le principe est de fixer un seuil de corrélation maximale S et d'éliminer des variables environnementales jusqu'à ce que toutes les corrélations soient plus petites que ce seuil. J'ai choisi un seuil $S = 0,9$ pour les modèles univariés, afin de filtrer les données les plus corrélées tout en conservant suffisamment de variables pour représenter les pressions de sélection possibles. Le choix de variables est effectué en trois étapes :

1. Calcul de la valeur absolue de la corrélation pour chaque paire de variables parmi les 72 variables environnementales.
2. Sélection de la plus haute corrélation $\sigma_{\max}$.
3. Si $\sigma_{\max}$ est supérieure au seuil S :
   — une des variables correspondantes est éliminée aléatoirement,
   — retour au point 2.

   Sinon, le processus s'arrête.

Le tirage aléatoire de la variable à éliminer (parmi les deux les plus corrélées) permet de répéter le processus pour obtenir des ensembles différents. L'application du processus permet d'aboutir à un jeu de 23 variables (voir table 4.5).

Comme le processus de sélection est en partie aléatoire, les ensembles obtenus ne contiennent pas forcément le même nombre de variables. Le jeu de données retenu est celui qui contient le plus de variables en fonction du seuil choisi et qui comprend en plus la longitude et la latitude. Les coordonnées ne sont pas des variables climatiques, mais je les ai incluses dans l'analyse de manière à faciliter l'interprétation des résultats au vu du gradient nord-sud des variables environnementales.

En sélectionnant les mêmes variables pour le jeu de données des 800k SNPs, la corrélation maximale entre les paires de variable est aussi inférieure à 0.9. J'ai donc utilisé les mêmes

Variable	Description
alt_SRTM	Altitude [m] (modèle : SRTM3)
aspect	Orientation
BIO2	Amplitude diurne moyenne (Moyenne des valeurs mensuelles de (temp. max - temp. min))
BIO3	Isothermalité (BIO2/BIO7) (· 100)
BIO7	Amplitude annuelle de la température (BIO5-BIO6)
BIO9	Température moyenne du trimestre le plus sec
BIO12	Précipitations annuelles
BIO15	Saisonnalité des précipitations (Coefficient de variation)
BIO18	Précipitations du trimestre le plus chaud
latitude	Latitude
longitude	Longitude
prec2	Précipitations en février
prec3	Précipitations en mars
prec4	Précipitations en avril
prec5	Précipitations en mai
prec6	Précipitations en juin
prec7	Précipitations en juillet
prec9	Précipitations en septembre
prec10	Précipitations en octobre
prec11	Précipitations en novembre
slope	Pente [%]
tmin10	Température minimale en octobre
tmax10	Température maximale en octobre

Table 4.5 – Liste des 23 variables environnementales utilisées pour les modèles univariés. Les températures sont indiquées en $[^{\circ}\mathrm{C} \cdot 10]$ et les précipitations en [mm].

variables environnementales pour les deux jeux de données moléculaires afin de faciliter la comparaison des résultats.

Pour les modèles corrélatifs bivariés

Limiter la multicollinéarité La sélection des variables explicatives pour un modèle multivarié doit prendre en compte leurs corrélations mutuelles. Si les prédicteurs sont trop interdépendants, les valeurs estimées pour les paramètres de la régression deviennent instables. Dans ce cas, ces paramètres peuvent varier radicalement en incluant ou excluant quelques points du modèle, alors qu'ils devraient rester stables lors d'une petite modification des points utilisés dans la régression. Ce phénomène de multicollinéarité peut être estimé en calculant le facteur d'inflation de la variance (*variance inflation factor*, VIF). Il dépend du coefficient de détermination R^2 de la régression linéaire d'un prédicteur en fonction des autres.

$$\mathrm{VIF} = \frac{1}{1 - R^2} \tag{4.2}$$

Le VIF est calculé pour chaque variable explicative et la valeur limite est généralement fixée à $\sim$5 (Dobson et Barnett, 2008).

Dans le cas des modèles bivariés, le VIF est également calculé avec une régression linéaire entre les prédicteurs. Le coefficient de détermination est alors égal au carré de la corrélation entre les variables. Un VIF plus petit que 5 correspond à une corrélation entre prédicteurs plus petite que 0,9. J'ai fixé la limite à 0,8, soit un VIF de $\sim$2,8 pour limiter la multicollinéarité et réduire le temps de calcul.

Variable d'appartenance D'autre part, afin d'évaluer l'efficacité des modèles multivariés dans la détection des signatures de sélection (question 2), j'ai inclus une variable supplémentaire dans l'analyse : « ankole » est le coefficient d'appartenance à la population ankole sur la base des résultats d'`Admixture` qui a servi à étudier la structure de population à la section 7.1.2. Cette variable sert à tester si une structure de population connue peut être prise en compte dans l'analyse.

Le processus de sélection du paragraphe précédent a été appliqué aux 24 variables, celles déjà choisies et « ankole », avec un seuil de corrélation de 0,8. Les 15 variables retenues sont résumées dans la table 4.6. Les corrélations entre les coordonnées géographiques et la variable « ankole » sont plus petites que le seuil fixé, c'est pourquoi la longitude et la latitude ont également été incluses dans l'analyse.

Dans un cadre de calcul intensif, la réduction du nombre de prédicteurs permet également de raccourcir le temps de traitement des modèles multivariés.

Variable	Description
ankole	Coefficient d'appartenance à la population ankole [%]
aspect	Orientation
BIO2	Amplitude diurne moyenne (Moyenne des valeurs mensuelles de (temp. max - temp. min))
BIO3	Isothermalité (BIO2/BIO7) ($\cdot$ 100)
BIO12	Précipitations annuelles
BIO15	Saisonnalité des précipitations (Coefficient de variation)
BIO18	Précipitations du trimestre le plus chaud
latitude	Latitude
longitude	Longitude
prec4	Précipitations en avril
prec5	Précipitations en mai
prec10	Précipitations en octobre
prec11	Précipitations en novembre
slope	Pente [%]
tmin10	Température minimale en octobre

Table 4.6 – Liste des 15 variables environnementales utilisées pour les modèles bivariés. Les températures sont indiquées en $[^{\circ}\mathrm{C} \cdot 10]$ et les précipitations en [mm].

4.6 Données simulées

La comparaison des résultats fournis par les différentes méthodes corrélatives de détection de signatures de sélection est confrontée à l'incertitude relative à l'influence réelle de la sélection sur les marqueurs analysés. L'utilisation de données simulées permet de tester le comportement de ces méthodes dans un cadre où les marqueurs neutres et adaptatifs sont clairement connus. J'ai comparé les résultats fournis par Samβada et `LFMM` pour un jeu de 100 SNPs simulés décrits par Jones et al. (2013) et que les auteurs m'ont gracieusement mis à disposition. Ces derniers ont utilisé `CDPOP`, qui est un simulateur de populations basé sur les individus. Ces données comprennent un loci soumis à la sélection et 99 loci neutres. L'objectif est de comparer la capacité de Samβada et de `LFMM` à distinguer les loci adaptatifs des loci neutres dans un contexte où ces caractéristiques sont connues [9].

`CDPOP` simule une population d'organismes dotés d'un patrimoine génétique. Les simulations sont organisées par itérations ; à chaque itération, tous les individus se déplacent et se reproduisent. Il y a deux types de simulations : dans le premier cas les générations d'individus sont séparées et les individus sont remplacés par leurs descendants à chaque itération, alors que dans le second cas les itérations représentent « une année », les générations se recouvrent temporellement et les individus vieillissent et meurent graduellement. Le patrimoine génétique est composé de loci bialléliques et la transmission des gènes suit l'hérédité mendélienne. L'utilisateur définit le nombre de locus, si certains d'entre eux sont soumis à la sélection (au

9. `CDPOP` considère que tous les individus appartiennent à une seule population, alors que `BayEnv` nécessite qu'ils soient répartis en populations. C'est pourquoi `BayEnv` n'a pas été utilisé dans ce cadre.

maximum deux) ainsi que le nombre d'allèles par loci au début de la simulation. L'utilisateur doit également fournir une liste des positions géographiques où les individus peuvent s'arrêter, ainsi que le coût d'un déplacement entre ces différents points, ce qui permet d'étudier l'influence des obstacles naturels sur l'évolution d'une population. La position des individus au début de la simulation peut être indiquée au logiciel ou tirée au hasard. A chaque itération, les organismes peuvent se déplacer pour trouver un partenaire. Le patrimoine génétique des nouveaux-nés est généré aléatoirement en fonction de celui de leurs parents, un nouvel allèle pouvant également apparaître en fonction du taux de mutation défini par l'utilisateur. L'effet de la sélection est appliqué à la naissance via un taux de mortalité lié au génotype. Dans le cas où un locus biallélique est adaptatif, trois génotypes sont possibles et la valeur sélective des individus est modélisée par trois surfaces de mortalité. Ces surfaces déterminent la probabilité qu'un individu meure à la naissance en fonction de son génotype et de sa position géographique.

Les données de Jones et al. simulent une population de 5'000 individus sur 1'000 générations séparées. Les individus ont la possibilité de parcourir jusqu'à 25% de la distance maximale possible pour trouver un partenaire. Le nombre de descendants suit une loi de Poisson de moyenne $\lambda = 4$. Ainsi tous les points de la grille sont occupés à chaque génération. Les individus excédentaires sont éliminés (par migration). Le patrimoine génétique comporte 99 loci neutres et un locus dominant soumis à la sélection. Les génotypes AA et AG ont la même valeur adaptative. La sélection s'applique en fonction de la latitude et le taux de mortalité varie linéairement entre 0 et sa valeur maximale. Trois scénarios différant par l'intensité de la sélection ont été répétés dix fois : *a*) La sélection d'intensité faible correspond à un taux de mortalité infantile maximal de 1%, *b*) l'intensité moyenne à un taux de 10% et *c*) la sélection forte à un taux de 50%. Les individus AA et AG ont un taux de mortalité nul au nord et maximal au sud, tandis que les individus GG ont un taux maximal au nord et nul au sud.

Après la description des données moléculaires et environnementales, le chapitre suivant est consacré au développement du logiciel proposé qui met en relation ces deux types d'information.

5 Développements bioinformatiques

La première étape du développement consiste à étudier les besoins et à comparer les avantages et inconvénients des approches possibles. Cette analyse m'a menée à la décision de développer un nouveau logiciel. La seconde partie du chapitre est consacrée à la conception et à l'implémentation de Samβada.

Afin d'atteindre le premier objectif de mon travail, j'ai considéré deux alternatives, soit étendre les fonctionnalités d'un programme existant (`MatSAM` voir p. 21), soit développer un nouveau programme.

5.1 `MatSAM`

5.1.1 Fonctionnalités disponibles

`MatSAM` est le premier logiciel de génomique environnementale développé dans le but de détecter des signatures de sélection. Il a été écrit par Stéphane Joost avec `MATLAB` [1] (Joost, 2006). `MatSAM` permet de modéliser la relation entre des marqueurs (recodés si nécessaire sous forme binaire), et des variables environnementales quantitatives par le moyen de régressions logistiques univariées. La significativité de chaque association est évaluée par les tests statistiques du rapport de vraisemblance et de Wald. Les résultats sont enregistrés sous forme de matrices et peuvent être analysés dans un tableur. `MatSAM` peut être utilisé de manière autonome (`.exe`) en installant la bibliothèque `MATLAB Compiler Runtime` [2]. Le logiciel propose également des macros destinées à faciliter la mise en évidence des modèles significatifs dans `Excel` [3] (Joost, Kalbermatten et Bonin, 2008).

`MatSAM` version 2, développé avec l'aide Michaël Kalbermatten, permet en plus d'utiliser des variables environnementales qualitatives et d'estimer la qualité d'ajustement (*goodness-of-fit*) avec plusieurs pseudo-R^2 (Joost et Kalbermatten, 2010). Le fichier de résultats inclut également

1. www.mathworks.fr
2. http://www.mathworks.fr/products/compiler/mcr/
3. www.microsoft.com

les macros nécessaires à Excel avec les paramètres corrects pour tester la significativité des modèles avec un seuil variable choisi par l'utilisateur.

Voici une rapide évaluation de MatSAM selon trois critères déterminants dans le cadre de la présente recherche.

5.1.2 Temps de calcul

Sur un ordinateur portable de type MacBookPro [4], MatSAM est capable de traiter un jeu de données comprenant 385 individus, 10 variables environnementales et 100 marqueurs binaires, soit 1'000 modèles, en 25 secondes. Le temps de calcul dépend de ces trois paramètres. Le nombre d'individus de l'exemple est du même ordre de grandeur que le nombre d'animaux échantillonnés dans le cadre de NextGen. Sur la base des données disponibles, le nombre de variables environnementales peut être estimé à une centaine au maximum. Le principal changement provient du séquençage intégral du génome qui fournit de volumineuses données moléculaires. Compte tenu du nombre de positions variables sur le génome des mammifères, nous attendions en un et dix millions de marqueurs génétiques. Si ces données sont des SNPs bialléliques, le recodage produit trois marqueurs binaires pour chaque loci, ce qui affecte également le temps de calcul.

En considérant 100 variables environnementales et un million de marqueurs, il y aurait au minimum 100 millions de modèles à analyser. Il faudrait donc environ 29 jours à MatSAM pour traiter les données de NextGen, peut-être dix fois plus suivant le nombre de marqueurs. Quant aux modèles bivariés, ils nécessiteraient environ 50 fois plus de temps.

MatSAM aurait besoin d'une refonte pour traiter de tels jeux de données.

5.1.3 Traitement des résultats

Le formatage des résultats en matrices permet un traitement aisé dans un tableur. De plus, l'inclusion de macros VisualBasic dans le fichier de résultats permet d'automatiser la sélection des modèles significatifs dans Excel. Cette méthode permet à l'utilisateur d'analyser ses résultats aisément et de créer ses propres graphiques. Il faut cependant relever que la bibliothèque de fonctions d'Excel est traduite dans chaque langue. Or MatSAM sauve les résultats avec les macros prérédigées en anglais. En conséquence, il est par exemple nécessaire de remplacer les "IF" par des "SI" pour analyser les résultats avec la version française d'Excel.

D'autre part, cette méthode ne permet pas de traiter de grands jeux de données. Les tableurs ne permettent généralement pas d'ouvrir des fichiers comprenant plusieurs millions de lignes, la limite d'Excel 2010 est d'un million, et le tri des résultats serait très long. De plus, l'analyse des résultats dans MATLAB directement nécessiterait de modifier la façon dont MatSAM stocke les modèles en mémoire.

4. MacBookPro5,3 : processeur Intel Core 2 Duo (2.8 GHz) et 8 Gb RAM.

5.1.4 Compatibilité

`MatSAM` requiert le `MATLAB Compiler Runtime` et il n'était pas sûr au début de l'année 2011 que cette bibliothèque serait disponible pour les ordinateurs 64 bits. La diffusion du programme auprès des utilisateurs aurait été compromise en cas de changement de la politique de `MathWorks`.

Compte tenu de la taille des données attendues, il est prévu de distribuer les calculs entre plusieurs ordinateurs. La plupart des clusters d'ordinateurs permettent généralement d'utiliser `MATLAB`. Mais, en revanche, rien ne garantit que les utilisateurs potentiels disposent de ce logiciel ou qu'ils aient les privilèges nécessaires pour installer la bibliothèque sur leur cluster.

Finalement, un objectif important étant de promouvoir la génomique environnementale et de rendre ces outils accessibles au plus grand nombre, en terme de facilité d'utilisation et d'adaptabilité au matériel disponible, j'étais réservée quant à laisser la distribution de cette nouvelle version de `MatSAM` dépendre d'un logiciel propriétaire.

5.1.5 Solution retenue

La poursuite du développement en `MATLAB` aurait nécessité un réusinage complet de `MatSAM`. N'ayant que peu d'expérience avec ce langage, j'ai choisi de développer une application autonome dédiée à la génomique environnementale. Cette solution a l'avantage de laisser une grande liberté d'implémentation et de faciliter la distribution du logiciel auprès des utilisateurs.

5.2 Développement d'un nouveau logiciel

5.2.1 Cahier des charges

Afin de remplir mes objectifs d'efficacité et d'accessibilité (p. 13), le nouveau logiciel doit répondre à trois exigences principales. Premièrement, il doit être **rapide** afin de permettre le calcul des régressions logistiques et le test de la significativité des modèles de manière efficace. Deuxièmement, il doit être **autonome** de manière à ce que son installation soit facile sur différents systèmes d'exploitation ainsi que sur les clusters. Et finalement, il doit octroyer à la communauté d'utilisateurs la **liberté** de le distribuer et de le modifier selon les besoins.

Le dernier critère est automatiquement rempli en développant un logiciel libre. En effet, une licence *open source* garantit que les utilisateurs puissent modifier le code selon leur besoins à la condition que le programme résultant soit *open source* lui également. Réciproquement, les éventuelles bibliothèques logicielles utilisées lors du développement doivent aussi être libres ou posséder une licence autorisant leur inclusion dans un logiciel *open source*.

5.2.2 Choix d'implémentation

Langage

Afin d'analyser efficacement de grands jeux de données, j'ai choisi d'utiliser un langage compilé, plus rapide qu'un langage interprété (comme R ou Python, Biggar et al., 2012). Les principales options étaient le langage Fortran que je connaissais assez bien, le C que je n'avais jamais utilisé et le C++ que je connaissais bien. Fortran est surtout utilisé en ingénierie et présente l'avantage de traiter directement des matrices. Son principal inconvénient est qu'il est globalement peu répandu et qu'ainsi peu d'utilisateurs auraient pu modifier un programme en Fortran. Le langage C est quant à lui couramment utilisé pour les systèmes embarqués, les systèmes d'exploitation et le calcul intensif, quand la vitesse d'exécution est importante. En contrepartie de sa rapidité, C est un langage de bas niveau, c'est-à-dire que les instructions élémentaires permettent des traitement très simples et que la création de fonctionnalités complexes exige un grand effort d'abstraction de la part de l'utilisateur. Finalement, C++ est un langage orienté objet très répandu. Son avantage est d'offrir un compromis entre performance et capacité d'abstraction. La bibliothèque standard (*C++ standard library*) inclut de nombreux conteneurs de données et algorithmes prédéfinis pour faciliter le développement d'applications. De plus, la plupart des bibliothèques de calcul numérique proposent une interface pour ce langage.

J'ai donc choisi de développer une application *open source* en C++.

Bibliothèques

Afin de faciliter le développement, j'ai cherché une bibliothèque pour le calcul matriciel en C++ contenant des fonctions statistiques, comme par exemple le calcul des quantiles d'une distribution. Plusieurs solutions ont été envisagées.

La bibliothèque `AlgLib` [5] répondait à ces exigences et comprenait également un module pour les régressions logistiques. La méthode de régression ne permettait toutefois pas d'inclure des variables qualitatives comme nous le souhaitions alors, ni de calibrer explicitement le modèle neutre.

Une autre option, la bibliothèque `LAPACK` [6], avait l'avantage d'être très répandue et d'être spécialisée dans les calculs matriciels. Le calcul des quantiles aurait cependant nécessité une autre bibliothèque et l'interface de `LAPACK` en C++ me paraissait difficile à utiliser.

J'ai finalement choisi la bibliothèque `Scythe Statistical Library` [7], qui propose une interface intuitive et possède de nombreuses fonctions statistiques (Pemstein et al., 2011). Cette bibliothèque *open source* est implémentée sous forme de *templates* C++, ce qui permet de

5. www.alglib.net
6. www.netlib.org/lapack/
7. scythe.wustl.edu

créer facilement de nouveaux types de matrices et rend `ScytheStat` très facile à intégrer dans un projet. Comme cette bibliothèque ne propose pas de module pour les régressions logistiques, j'ai implémenté directement la méthode de Dobson et Barnett (2008) pour l'ajustement de modèles logistiques.

5.2.3 Genèse de Samβada

La première version de Samβada, nommée `ScytheSAM` durant le développement, calculait une régression logistique univariée pour chaque couple marqueur-variable environnementale. Toutes les variables explicatives devaient alors être continues. Le coeur du programme n'a pas changé dans les versions ultérieures : la régression logistique utilise la méthode du maximum de vraisemblance et la significativité des modèles est déterminée avec les tests de Wald et du rapport de vraisemblance (G) (Dobson et Barnett, 2008, et sec. 6.1.1).

A ce stade, j'ai étudié la possibilité d'inclure des variables environnementales discrètes. Cette fonctionnalité offerte par `MatSAM 2` permet d'utiliser des variables nominales et ordinales (Joost et al., 2012). La gestion des variables discrètes semblait pourtant difficile à combiner avec les modèles multivariés. En effet, une variable quantitative doit être recodée en plusieurs variables indicatrices, ce qui implique que les modèles comprenant le même nombre de variables environnementales n'ont pas forcément le même nombre de variables prédictrices lors de l'analyse. Cette caractéristique complique le codage des variables environnementales en mémoire ainsi que la sauvegarde des coefficients de régression, car il peut y avoir plusieurs coefficients par variable. Comme l'implémentation des modèles multivariés quantitatifs était prioritaire, j'ai concentré le développement sur les variables continues.

Lors d'une visite académique dans notre laboratoire, avec un post-doctorant égyptien nous avons mené au printemps 2011 une étude sur les moutons qui comprenait 1'483 individus et 49'034 SNPs (Agha, 2011 ; Stucki et al., 2012). Cette recherche a révélé un problème latent : certains modèles incluant de nombreux individus peuvent avoir des p-valeurs extrêmement petites. Elles sont parfois si petites que l'ordinateur considère qu'elles sont égales à zéro et les modèles très significatifs ne sont alors plus distinguables. En parallèle, l'augmentation du nombre de modèles analysés abaisse le seuil de significativité quand on applique la correction de Bonferroni. Le risque est donc d'obtenir des seuils de significativité frôlant le plus petit nombre représentable en mémoire et des modèles significatifs avec des p-valeurs nulles.

Pour éviter cet écueil, j'ai décidé que Samβada ne se baserait pas sur les p-valeurs associées aux modèles. Le seuil de significativité choisi par l'utilisateur est directement converti en un seuil de score pour les deux tests statistiques utilisés. Un modèle est significatif si ses scores G et Wald sont supérieurs au seuil. De la sorte, les modèles non significatifs peuvent être effacés directement ou conservé dans les résultats.

Les principaux modules de Samβada sont *a*) l'analyse des modèles univariés (5.3.1) ; *b*) l'analyse des modèles multivariés (5.3.2) ; *c*) la mesure de l'autocorrélation spatiale (5.3.3) ; *d*) la

version dédiée au calcul haute performance (5.3.4) ; *e*) la possibilité de distribuer les calculs sur plusieurs machines ; et *f*) la visualisation des résultats (5.3.5).

5.3 Détails de l'implémentation

L'objectif de Samβada est d'appliquer des régressions logistiques à de grands jeux de données et de les résoudre très rapidement, tout en étant fiable et en donnant également la possibilité de traiter des modèles multivariés, et en fournissant des statistiques spatiales en complément. Ces exigences cumulées ne sont pas forcément compatibles. Par exemple la gestion des valeurs manquantes lors du calcul de l'autocorrélation spatiale nécessite de vérifier la pondération des points selon la distance les séparant avant d'analyser chaque variable et de la corriger au besoin, et cela consomme des ressources de calcul et donc du temps.

De plus, les tests et les premiers commentaires des utilisateurs (Kevin Leempoel, Stéphane Joost, Ivo Widmer et les étudiants du cours de *Geocomputation*) ont également montré que les deux types d'utilisation seraient distincts : *a*) les grands jeux de données seraient avant tout traités avec des régressions logistiques univariées ; *b*) les analyses multivariée et spatiale seraient réalisées dans un second temps sur une sélection de marqueurs intéressants. Comme les jeux de données comprenant quelques dizaines de milliers de marqueurs peuvent être traités sur un seul ordinateur en une nuit, la plupart des utilisateurs effectueraient donc toutes les analyses sur la même machine.

Enfin Samβada utilise une bibliothèque externe pour inverser des matrices de dimension quelconque lors de l'ajustement des régressions logistiques et également pour calculer la fonction de répartition de la loi du χ^2 lors des tests de significativité des modèles. Or ces traitements sont simplifiés dans le cas des modèles univariés (en effet, les matrices à inverser sont de taille 2×2, l'inverse est donc donnée par une formule, et la loi du χ^2 à un degré de liberté revêt une forme particulière qui facilite son calcul). Le traitement des régressions logistiques univariées ne nécessite donc pas de recourir à une bibliothèque externe.

C'est pourquoi j'ai scindé le développement de Samβada entre une version Desktop centrée sur les modèles multivariés, sur l'autocorrélation spatiale et sur la flexibilité d'utilisation, alors que le traitement de grands jeux de données est réalisé par `CoreSAM`, écrit en C et dédié à la calibration des régressions logistiques univariées. La simplification du processus permet de gagner en efficacité. Les logiciels sont disponibles à l'adresse lasig.epfl.ch/sambada.

5.3.1 Samβada Desktop et modèles univariés

Samβada est un logiciel autonome qui ne nécessite pas d'installer d'autres programmes au préalable. Il fonctionne en ligne de commande et prend en arguments les noms d'un fichier de paramètres et de deux fichiers de données. Ces derniers sont, d'une part, les marqueurs génétiques, qui doivent être binaires puisque la régression logistique modélise la présence

ou l'absence d'un allèle ; et d'autre part les variables environnementales. Ces fichiers de données doivent contenir un individu par ligne et une variable par colonne. Les données environnementales et génétiques peuvent être dans le même fichier ou dans des fichiers séparés. Dans ce cas, les individus doivent impérativement apparaître dans le même ordre. Les colonnes reçoivent un nom par défaut si l'utilisateur n'en a pas fourni, ce qui permet de reconnaître les modèles. Les fichiers peuvent contenir des colonnes qui ne doivent pas être analysées, comme la race ou le nom de l'éleveur. Il suffit de les désigner comme « facultatives » dans les paramètres. Les caractères de fin de ligne sont reconnus automatiquement et le caractère séparant les colonnes peut être précisé dans le fichier de paramètres. La figure 5.1 présente un exemple de fichier de paramètres pour Samβada et la figure 5.2 est un extrait des résultats pour les modèles univariés. La structure du programme est présentée à l'annexe B.

```
    HEADERS YES
    WORDDELIM " "
*   NUMVARENV 24
*   NUMMARK 120103
*   NUMINDIV 804
    IDINDIV short_name ID_indiv
    SPATIAL longitude latitude SPHERICAL NEAREST 20
    AUTOCORR BOTH MARK 1000
*   DIMMAX 1
*   SAVETYPE END BEST 0.01
```

Figure 5.1 – Exemple de fichier de paramètres pour l'analyse avec Samβada. Chaque ligne contient une option de calcul, celles qui sont signalées par une croix dans la marge sont obligatoires. L'ordre des lignes n'a pas d'importance. Dans cet exemple, les deux premières lignes indiquent que les fichiers de données contiennent une ligne d'en-tête et que les valeurs sont séparées par des espaces. Puis viennent le nombre de variables environnementales, le nombre de marqueurs et le nombre d'individus/échantillons. L'option `IDINDIV` indique les noms des colonnes contenant les identifiants des individus ; dans cet exemple les données génétiques et environnementales sont enregistrées dans des fichiers séparés. Les deux lignes suivantes concernent la mesure de l'autocorrélation spatiale, avec les noms des coordonnées, qui sont de type sphérique, le schéma de pondération et la bande passante ; ici les 20 plus proches voisins sont considérés. L'analyse inclura l'autocorrélation globale et locale (`BOTH`) des marqueurs génétiques (`MARK`) et la significativité sera évaluée avec 1'000 permutations. L'option suivante signifie que la détection de la sélection se basera sur des modèles univariés (`DIMMAX 1`). La dernière ligne indique que les résultats seront sauvés à la fin des calculs, que seuls les modèles significatifs ayant un parent significatif seront enregistrés et que le seuil de significativité est fixé à 1% (avant la correction de Bonferroni).

Modèles multivariés

Pour chaque marqueur, Samβada calcule un modèle univarié par paramètre environnemental. Puis il considère toutes les combinaisons possibles de variables pour les modèles bivariés, trivariés, etc. jusqu'à la « dimension » (nombre de variables environnementales) maximale

Marker	Env_1	Loglikelihood	Gscore	WaldScore	NumError	Efron	McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AIC	BIC	Beta_0	Beta_1
Hapmap41074-BTA-73520_AA	prec7	-443.11	208.53	151.72	0	0.25	0.19	0.19	0.23	0.10	890.22	912.98	-2.04	0.03
ARS-BFGL-NGS-113888_GG	prec7	-441.73	208.67	151.70	0	0.25	0.19	0.19	0.23	0.10	887.47	910.23	-2.02	0.03
Hapmap41762-BTA-117570_GG	prec7	-435.96	202.93	148.43	0	0.24	0.19	0.19	0.22	0.10	875.92	898.68	-1.86	0.03
ARS-BFGL-NGS-46098_GG	prec7	-440.04	200.82	147.60	0	0.24	0.19	0.18	0.22	0.10	884.07	906.83	-1.88	0.03
ARS-BFGL-NGS-113888_GG	latitude	-449.13	193.89	146.89	0	0.23	0.18	0.17	0.21	0.09	902.25	925.01	-0.73	0.86
Hapmap41074-BTA-73520_AA	latitude	-450.81	193.13	146.61	0	0.23	0.18	0.17	0.21	0.09	905.62	928.38	-0.75	0.85
Hapmap41762-BTA-117570_GG	latitude	-444.40	186.04	141.99	0	0.21	0.17	0.17	0.21	0.09	892.80	915.56	-0.57	0.84
ARS-BFGL-NGS-113888_GG	prec6	-455.48	181.19	138.85	0	0.21	0.17	0.16	0.20	0.09	914.95	937.71	-2.22	0.03
Hapmap41074-BTA-73520_AA	prec6	-457.38	179.99	138.13	0	0.21	0.16	0.16	0.20	0.09	918.77	941.53	-2.23	0.03
ARS-BFGL-NGS-46098_GG	latitude	-451.22	178.45	138.11	0	0.21	0.17	0.16	0.20	0.09	906.44	929.20	-0.59	0.82
Hapmap41813-BTA-27442_AA	prec7	-462.30	179.89	137.52	0	0.22	0.16	0.16	0.20	0.08	928.60	951.36	-1.92	0.03
ARS-BFGL-NGS-46098_GG	prec6	-451.51	177.87	137.27	0	0.21	0.16	0.16	0.20	0.09	907.03	929.78	-2.11	0.03
BTA-73516-no-rs_AA	prec7	-460.18	177.43	136.04	0	0.21	0.16	0.16	0.20	0.08	924.35	947.11	-1.83	0.03
Hapmap41813-BTA-27442_AA	latitude	-469.89	164.71	130.98	0	0.20	0.15	0.15	0.19	0.08	943.77	966.53	-0.76	0.76
Hapmap41762-BTA-117570_GG	prec6	-454.17	166.51	130.97	0	0.20	0.15	0.15	0.19	0.08	912.33	935.09	-1.96	0.03
ARS-BFGL-NGS-46098_GG	longitude	-458.86	163.18	130.95	0	0.18	0.15	0.15	0.18	0.08	921.72	944.48	-23.95	0.76
Hapmap41074-BTA-73520_AA	bio7	-457.07	180.61	129.73	0	0.21	0.16	0.16	0.20	0.09	918.14	940.90	-11.85	0.08
ARS-BFGL-NGS-113888_GG	bio7	-456.32	179.50	128.90	0	0.20	0.16	0.16	0.20	0.09	916.64	939.40	-11.82	0.08
BTA-73516-no-rs_AA	latitude	-468.36	161.06	128.61	0	0.19	0.15	0.14	0.18	0.08	940.72	963.48	-0.67	0.76
Hapmap28985-BTA-73836_CC	prec6	-457.78	157.45	125.68	0	0.19	0.15	0.14	0.18	0.08	919.57	942.33	1.87	-0.03
Hapmap31863-BTA-27454_GG	prec7	-474.85	155.28	123.46	0	0.19	0.14	0.14	0.18	0.07	953.70	976.43	-1.91	0.02
ARS-BFGL-NGS-46098_GG	bio7	-456.70	167.50	121.71	0	0.20	0.15	0.15	0.19	0.08	917.39	940.15	-11.35	0.08
BTA-73516-no-rs_AA	prec6	-474.90	147.99	119.50	0	0.17	0.13	0.13	0.17	0.07	953.79	976.55	-1.97	0.03
Hapmap41762-BTA-117570_GG	bio7	-460.77	153.30	113.69	0	0.18	0.14	0.14	0.17	0.07	925.54	948.30	-10.71	0.07
Hapmap28985-BTA-73836_GG	bio3	-381.27	160.94	111.21	0	0.21	0.17	0.17	0.18	0.10	766.54	789.30	19.98	-0.26
ARS-BFGL-NGS-113888_GG	bio3	-471.77	148.61	106.51	0	0.17	0.14	0.13	0.17	0.07	947.53	970.29	20.21	-0.24

Figure 5.2 – Extrait d'un fichier de résultats de Samβada pour des modèles univariés, il y a un modèle par ligne. La première colonne est le nom du marqueur génétique, ici le nom du loci combiné au nom de l'allèle. La deuxième colonne est le nom de la variable environnementale. Puis viennent la log-vraisemblance, le score G, le score de Wald et le code d'erreur (0 en cas de réussite). Les cinq colonnes suivantes sont des mesures d'ajustement de la régression (pseudo-R^2). L'analyse comprend également les critères AIC (*Akaike information criterion*) et BIC (*Bayesian information criterion*). Les deux dernières colonnes sont les paramètres β de la régression, à savoir le paramètre constant et celui se rapportant à la variable environnementale. Les fichiers de résultats pour les modèles multivariés comportent des colonnes supplémentaires pour les variables environnementales (Env_2, Env_3, ...) et pour les paramètres de régression (Beta_2, Beta_3, ...).

choisie par l'utilisateur. Les scores G et Wald sont calculés pour chaque modèle ainsi que les critères AIC et BIC et que cinq « pseudos-R^2 » qui mesurent la qualité de l'ajustement (*goodness-of-fit*). A la fin du calcul, les modèles sont triés selon leur score de Wald (le plus conservateur). Samβada crée un fichier de résultats par dimension et écrit un modèle par ligne dans le fichier de résultats. Dans le cas où il y aurait trop de modèles à calculer par rapport à la mémoire disponible, un paramètre permet d'enregistrer les résultats sur le disque dur pendant le calcul. Ils ne sont alors pas triés.

5.3.2 Algorithme spécifique pour les modèles multivariés

La régression logistique se généralise aisément au cas à plusieurs variables. Le calcul d'un modèle suit la même procédure que dans le cas univarié, seule la taille des matrices change. Samβada analyse l'effet cumulé d'une combinaison de prédicteurs avec un modèle additif : les interactions entre prédicteurs ne sont pas considérées. L'objectif est de déterminer quelle combinaison de variables environnementales explique le mieux, et le plus parcimonieusement, la distribution de chaque marqueur génétique. Dans la suite du texte, la « dimension » d'un modèle désigne le nombre de variables environnementales qu'il comprend.

Ensemble des modèles possibles

L'analyse multivariée nécessite de déterminer l'ensemble des modèles possibles pour une dimension donnée. Chaque marqueur est indépendant. Considérons un marqueur et cinq variables environnementales, E1,..., E5.

— En dimension 1, il y a un modèle (univarié) par couple marqueur-prédicteur.

— En dimension 2, chaque paire de variables ne doit être considérée qu'une fois ({E1, E2} = {E2, E1}).

— En dimension 3, le groupe {E1, E2, E3} peut être ordonné de 6 manières différentes ({E1, E2, E3} = {E2, E3, E1} = {E3, E1, E2} = {E1, E3, E2} = {E3, E2, E1} = {E2, E1, E3}).

En effet, le nombre de combinaisons de q variables environnementales parmi k est $C_q^k = \frac{k!}{q!(k-q)!}$ (voir p. ex. Morgenthaler, 2007). La méthode la plus simple pour parcourir de toutes les combinaisons suit un schéma récursif de « déplacement de pions » résumé sur la figure 5.3. Pour les modèles à q dimensions, q pions indiquent les variables incluses dans le modèle courant. Les pions sont placés sur les q premières variables. Le pion de droite est déplacé successivement d'un cran vers la droite jusqu'à la dernière variable (Fig. 5.3, points 1 à 4). Le pion situé à sa gauche est alors déplacé d'un cran vers la droite, tandis que ce pion est déplacé sur la variable à droite du précédent (point 5) et le mouvement reprend (points 6 à 10). Quand le deuxième pion a atteint l'avant dernière case, le troisième est alors déplacé d'un cran, etc. Toutes les combinaisons ont été considérées lorsque les pions occupent les q cases de droite.

Les déplacements peuvent être vus comme des boucles imbriquées. La figure 5.4 présente le parcours de tous les modèles à deux et trois variables parmi k.

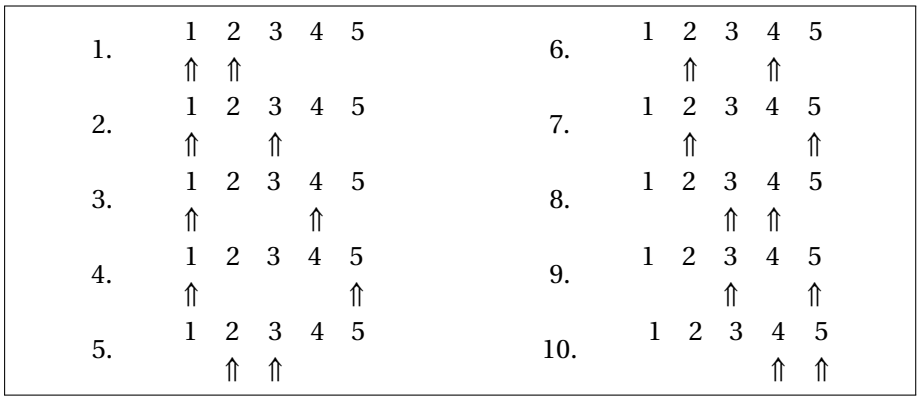


Figure 5.3 – Parcours de tous les modèles bivariés avec $k = 5$ variables environnementales. Les étapes sont numérotées de 1 à 10. A chaque étape, les flèches indiquent quelles variables environnementales sont incluses dans le modèle.

```
for i = 1 to k − 1 do
   for j = i + 1 to k do
      Calcul du modèle {E_i, E_j}
   end for
end for
```

(a) Modèles bivariés

```
for i = 1 to k − 2 do
   for j = i + 1 to k − 1 do
      for l = j + 1 to k do
         Calcul du modèle {E_i, E_j, E_l}
      end for
   end for
end for
```

(b) Modèles trivariés

Figure 5.4 – Algorithme pour parcourir tous les modèles à deux et trois variables avec des boucles imbriquées.

Cette méthode récursive a deux inconvénients. D'une part, chaque dimension de modèle implique un nombre de boucles différent et doit donc doit être traitée séparément, ce qui duplique le code et le rend difficile à maintenir. D'autre part, comme chaque dimension doit être traitée spécifiquement, cette approche nécessite de déterminer pendant le développement quelle sera la dimension maximale que l'utilisateur pourra choisir pour ses modèles.

Algorithme de la génération précédente

Il faut donc trouver une méthode qui permette de parcourir tous les modèles, quelle que soit leur dimension, avec un seul algorithme afin d'éviter de dupliquer le code source. J'ai considéré plusieurs solutions, la plus simple (et efficace) part de l'observation que les modèles à q variables peuvent être construits à partir des modèles à $q-1$ variables. Cette solution est appelée « algorithme de la génération précédente » dans la suite du texte.

Sur les points 1 à 4 de la fig. 5.3, le pion de gauche reste sur la variable E1 alors que le pion de droite parcourt les variables restantes, de E2 à E5. Puis le pion de gauche est déplacée d'un cran pendant que celui de droite parcourt les variables E3 à E5 (points 5 à 7). Ensuite le premier pion marque la variable E3, tandis que le second indique tour à tour E4 et E5, et enfin le dernier modèle implique les variables E4 et E5. En d'autres termes, le pion de gauche a parcouru les variables E1 à E4, de la même manière que dans le cas des modèles univariés. Pour chaque position occupée par ce pion, le second pion a parcouru toutes les variables à sa droite. En résumé, le parcours de tous les modèles bivariés peut être vu comme la combinaison de deux mouvements : le pion de gauche a parcouru les « modèles univariés » allant de E1 à E4 tandis que le pion de droite a sélectionné tour à tour les variables libres à sa droite. Le pion de gauche n'a pas pas été placé sur la variable E5, car il n'y avait plus de variables libres à sa droite pour placer l'autre pion. Les modèles bivariés impliquant la variable E5 ont tous été considérés précédemment (fig. 5.3, étapes 4, 7, 9 et 10). Dans cet exemple, le pion de gauche représente le fait qu'on parcourt tous les modèles de la génération précédente, pendant que le dernier pion indique successivement chaque variable disponible « à droite » de celles déjà utilisées. L'avantage de cette méthode qui construit chaque génération de modèles à partir de la précédente est qu'elle est indépendante du nombre total de variables considérées.

Pour construire les modèles à q variables, l'algorithme de la génération précédente parcourt tous les modèles à $q-1$ variables. Pour chacun d'entre eux, il place le dernier pion tour à tour sur chaque variable disponible ayant un numéro supérieur aux variables déjà sélectionnées. Dans l'exemple de la fig. 5.3, les variables disponibles sont situées « à droite » des premières. Cette règle pour le dernier pion garantit que chaque combinaison de q variables sera considérée une seule fois. J'appelle les modèles de dimension $q-1$ la « génération précédente », car c'est celle qui sert à générer les modèles de dimension q. Chaque modèle de dimension $q-1$ permet donc de créer des modèles de dimension q en lui ajoutant une variable, comme résumé sur la fig. 5.5. Les principaux avantages de cet algorithme par générations sont les règles simples de déplacement de pions et le fait de pouvoir construire des modèles de n'importe quelle dimension.

Significativité des modèles multivariés

Après avoir développé une méthode pour construire l'ensemble des modèles multivariés, il s'agit de calculer leur significativité. Pour le test du rapport de vraisemblance (G), chaque modèle multivarié pourrait être comparé au modèle constant incluant uniquement la fréquence moyenne du marqueur. Cette approche ne permet pas de déterminer si toutes les variables du modèle sont importantes ou si certaines pourraient être écartées. C'est pourquoi chaque modèle multivarié est comparé à des modèles plus simples de la génération précédente. Je définis les *parents* d'un modèle M à q variables comme étant les modèles de dimension $q-1$ obtenus en écartant une variable de M. Les parents de M sont les modèles à partir desquels M peut être généré. Dans la fig. 5.3, le deuxième modèle M$\{E_1, E_3\}$ a pour parents M$\{E_1\}$ et M$\{E_3\}$. Le modèle M$\{E_1\}$ est facile à retrouver en mémoire car il a servi à créer M$\{E_1, E_3\}$. Il faut donc

if q=1 **then** // cas univarié
  for $i = 1$ **to** k **do**
    Calcul du modèle $\{E_i\}$
  end for
else // cas multivarié
  for all modèles M $\in$ « génération précédente » **do**
    Soit E_c la dernière variable de M
    if $c < k$ **then** // test si E_c n'est pas la dernière variable possible E_k
      for $i = c+1$ **to** k **do** // parcours des variables disponibles « à droite » de E_c
        Calcul du modèle $\{M, E_i\}$
      end for
    end if
  end for
end if

Figure 5.5 – Algorithme par génération pour construire les modèles à q variables à partir de la génération précédente (modèles à $q-1$ variables). Les variables environnementales sont numérotées E_1 à E_k.

retrouver M$\{E_3\}$.

J'ai opté pour une solution basée sur la « bibliothèque standard » de C++ (*C++ standard library*). C++ définit des « paires » et des « ensembles » d'objets (Stroustrup, 2013). Un modèle est désigné sans équivoque par son marqueur et ses variables environnementales. L'étiquette d'un modèle est donc une paire formée d'un numéro de marqueur et d'un ensemble de numéros de variables environnementales.

```
typedef pair< int, set< int > > etiquetteModele;
```

Les `pair` et les `set` possèdent une relation d'ordre (définie par le compilateur) qui rend la recherche d'un élément efficace [8]. De plus, le temps nécessaire pour effacer un élément dans un set est constant (amorti). Ainsi, pour rechercher un parent de M, il suffit de copier l'étiquette de M, de supprimer une des variables environnementales et de rechercher le modèle de dimension $q-1$ correspondant à cette nouvelle étiquette. Les parents de M sont réunis en répétant cette opération pour chaque variable environnementale de M. Le test du rapport de vraisemblance qui est basé sur cette approche est présenté à la sec. 6.1.1.

8. La version du compilateur `gcc` que j'utilise n'inclut pas les `unordered_set` de la norme C++11 (ISO/IEC JTC1/SC22/WG21, 2011 ; Stroustrup, 2013). J'ai sacrifié un peu d'optimisation pour garder un programme compatible avec la plupart des compilateurs.

5.3.3 Autocorrélation spatiale

L'autocorrélation spatiale est le phénomène par lequel des points proches ont tendance à se ressembler ou à différer plus que ce à quoi on pourrait s'attendre si ces points étaient distribués au hasard dans l'espace. La mesure de l'autocorrélation spatiale dans un jeu de données permet de déterminer si les points peuvent être considérés comme indépendants. Comme cette hypothèse est à la base de nombreux tests statistiques, il est important d'évaluer sa validité. Samβada mesure l'autocorrélation spatiale des marqueurs moléculaires et des variables environnementales en calculant le I de Moran (Moran, 1950, et sec. 6.1.2) global et local (*Local Indicators of Spatial Association,* Anselin, 1995). Samβada propose quatre schémas de pondération spatiale pour déterminer quels points sont considérés comme voisins (et dans quelles proportions). La significativité de la mesure est testée en permutant les valeurs des points. Les résultats peuvent être sauvés au format texte ou *shapefile* pour être importés dans un logiciel SIG. Le fonctionnement de ce module est décrit en détail à la section 6.1.2.

5.3.4 Calcul distribué (`CoreSAM`)

Lorsque les données deviennent trop volumineuses pour être traitées en un temps raisonnable sur un ordinateur, les calculs peuvent être accélérés en les répartissant entre plusieurs machines. Samβada et `CoreSAM` utilisent le parallélisme des données car chaque modèle peut être traité indépendamment. Les variables moléculaires représentent la quasi-totalité des données à transférer, elles sont donc découpées en segments pour être distribuées. Les variables environnementales sont très légères et sont copiées sur chaque machine. En comparaison d'un calcul effectué sur un seul ordinateur, le fichier de paramètres transmis à chaque machine contient en plus le nombre total de marqueurs moléculaires pour calculer la correction de Bonferroni.

Samβada et `CoreSAM` gèrent le calcul distribué en trois étapes grâce au module `Supervision` :

Division `Supervision` découpe les données moléculaires en segments. Chaque fichier est renommé automatiquement en incluant un numéro d'identification et le numéro du premier marqueur du lot. Cette information est nécessaire pour nommer les variables si les données n'ont pas d'en-tête.

Calcul Chaque machine reçoit le fichier de paramètres, les données environnementales et son lot de marqueurs génétiques à traiter. Le calcul se déroule de la même manière qu'avec un seul processus.

Fusion Les résultats sont réunis sur une machine. Il y a un fichier par lot et un par dimension que le module `Supervision` reconnaît en fonction du nom. Les résultats correspondant à chaque dimension sont fusionnés et triés. `Supervision` permet de choisir le score utilisé pour le tri et de fixer un score minimal en deça duquel les résultats ne sont pas copiés. Cette option permet d'éliminer des modèles ayant un score très faible afin d'accélérer la sélection des modèles significatifs si les résultats sont analysés avec un logiciel externe (p. ex. `R`). `Supervision` exporte les résultats dans un fichier par

dimension, exactement comme s'ils avaient été traités d'un bloc.

5.3.5 Visualisation des résultats

Comme Samβada et CoreSAM fonctionnent en ligne de commande, les résultats sont représentés graphiquement par d'autres programmes. La visualisation des résultats peut être effectuée dans R grâce à des scripts fournis avec Samβada qui permettent les analyses suivantes :

a) La sélection simultanée de modèles en fonction de plusieurs seuils de significativité ;

b) L'identification des loci correspondants à un groupe de modèles ;

c) La représentation graphique de la distribution des p-valeurs des modèles associés à une variable environnementale ;

d) La visualisation de chaque chromosome et de ses loci soumis à la sélection. Ceux-ci sont groupés en fonction des distances en paires de bases les séparant, ce qui permet de comparer la densité de marqueurs entre les régions.

e) Les indices d'autocorrélation spatiale peuvent être importés dans tous les logiciels SIG, par exemple dans QuantumGIS.

Les résultats peuvent donc être analysés avec d'autres logiciels libres.

6 Méthodes statistiques pour détecter la sélection naturelle

Ce chapitre présente les modèles mathématiques sous-tendant l'analyse de Samβada. L'accent est mis sur l'évaluation de la significativité des modèles multivariés et sur la mesure de l'autocorrélation spatiale. La fin du chapitre est consacrée à trois approches de détection des signatures de sélection naturelle et à une approche pour identifier la structure d'une population qui seront utilisées au chapitre suivant.

6.1 Bases de Samβada

6.1.1 Méthodes corrélatives

Les méthodes corrélatives en génomique environnementale modélisent la probabilité d'occurrence de marqueurs génétiques en fonction des caractéristiques des habitats des individus. En effet, si un allèle fournit un avantage dans un environnement, les individus qui le portent auront en moyenne plus de descendants que les autres. Lors de ce processus d'adaptation, l'allèle favorable devient plus fréquent dans cet environnement. Ainsi, les marqueurs génétiques soumis à la sélection naturelle présentent des fréquences alléliques variables entre les différents habitats. Les méthodes corrélatives utilisent le plus souvent des modèles linéaires généralisés pour relier les fréquences alléliques aux conditions environnementales ; les associations significatives entre génome et environnement indiquent alors quels marqueurs sont potentiellement soumis à la sélection naturelle.

Samβada utilise une approche basée sur les individus, c'est-à-dire qu'il modélise la probabilité d'occurrence d'un allèle chez chaque individu en fonction de la composition environnementale de son habitat. D'autres approches comme `BayEnv` modélisent la fréquence allélique d'un marqueur chez un groupe d'individus partageant le même environnement.

Modèles linéaires généralisés

Les modèles linéaires permettent de représenter des phénomènes où la variable observée est proportionnelle à un ou plusieurs facteurs explicatifs :

$$\begin{aligned} E(Y_i) &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ Y_i &\sim N(\mu_i, \sigma_i) \end{aligned} \qquad \forall i \in \{1, \ldots, n\} \text{ observations} \tag{6.1}$$

Ils partagent leurs propriétés avec une classe plus étendue de modèles qui permettent de décrire une grande variété de phénomènes où l'observation n'est pas directement proportionnelle aux facteurs. Ces derniers sont les *modèles linéaires généralisés* (*generalised linear models*, GLM, Dobson et Barnett, 2008), de la forme :

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{6.2}$$

$$\mu_i = E(Y_i) \tag{6.3}$$

$$f(y_i; \theta_i) = \exp\left[y_i b(\theta_i) + c(\theta_i) + d(\theta_i)\right] \tag{6.4}$$

L'éq. 6.2 décrit la relation entre les variables observées $\boldsymbol{x}_i$, les paramètres à estimer $\boldsymbol{\beta}$ et μ_i, l'espérance des mesures au point i (éq. 6.3). La *fonction de lien* g (*link function*) est monotone et différentiable. Les observations $Y_1, \ldots, Y_n$ suivent une distribution exponentielle canonique et sont toutes de la même forme f (éq. 6.4). Par conséquent, la densité de probabilité conjointe des Y_i peut s'écrire :

$$f(y_i, \ldots, y_n; \theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} \exp\left[y_i b(\theta_i) + c(\theta_i) + d(\theta_i)\right] \tag{6.5}$$

$$= \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(\theta_i)\right] \tag{6.6}$$

Les fonctions b, c et d sont dérivables, θ_i est un paramètre propre à Y_i et $\mu_i = E(Y_i)$ est une fonction de θ_i. Comme les θ_i peuvent différer pour chaque observation, les modèles se concentrent sur un ensemble réduit de paramètres $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T (p < N)$.

Les variables explicatives $\boldsymbol{x_i}$ sont des vecteurs $p \times 1$

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} \quad \text{ou} \quad \boldsymbol{x}_i^T = \begin{bmatrix} x_{i1} \ \ldots \ x_{ip} \end{bmatrix} \tag{6.7}$$

et $\boldsymbol{\beta}$ est le vecteur $p \times 1$ des paramètres :

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{6.8}$$

Le vecteur $\boldsymbol{x}_i^T$ est la i^e ligne de la matrice $\boldsymbol{X}$ qui contient une ligne par observation et une colonne par variable explicative.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{N1} & \dots & x_{Np} \end{bmatrix} \tag{6.9}$$

Le choix des fonctions b, c et d détermine la loi de probabilité. La distribution binomiale s'obtient selon l'éq. 6.10 où le paramètre θ_i est généralement noté π_i et n_i désigne le nombre d'événements considérés pour l'observation i.

$$\begin{aligned} b(\pi_i) &= \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ c(\pi_i) &= n\ln(1-\pi_i) \\ d(\pi_i) &= C_{y_i}^{n_i} = \frac{n_i!}{y_i!(n_i-y_i)!} \\ \Rightarrow f(y_i;\pi_i) &= \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\ln(1-\pi_i) + \ln\left(C_{y_i}^{n_i}\right)\right] \qquad (6.10) \\ &= \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} \cdot (1-\pi_i)^{n_i} \cdot C_{y_i}^{n_i} \\ &= C_{y_i}^{n_i} \cdot \pi_i^{y_i} \cdot (1-\pi_i)^{n_i-y_i} \qquad (6.11) \end{aligned}$$

Modèle logistique

Le modèle logistique est basé sur la loi binomiale. Il tire son nom de la fonction de lien *logit* :

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q \tag{6.12}$$

$$\pi_i = \frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)} \tag{6.13}$$

Cette fonction garantit que la probabilité d'occurrence soit comprise entre 0 et 1 quelles que soient les valeurs de $\boldsymbol{x_i}$ et $\boldsymbol{\beta}$. Si on considère n points de mesure, où un évenément peut se produire n_i fois avec une probabilité π_i pour chaque point i, la fonction de distribution conjointe s'obtient en combinant les éq. 6.5, 6.10 et 6.12 :

$$f(y_i,\dots,y_n;\pi_1,\dots,\pi_n) = \exp\left[\sum_{i=1}^{n}\left(y_i\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) - n_i\ln\left(1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right) + \ln C_{y_i}^{n_i}\right)\right] \tag{6.14}$$

La calibration du modèle consiste à trouver la valeur optimale de $\boldsymbol{\beta}$ pour que les probabilités π_i décrivent précisément les fréquences observées $\frac{y_i}{n_i}$.

Méthode du maximum de vraisemblance

La calibration de modèles linéaires généralisés utilise couramment la méthode du maximum de vraisemblance. La fonction de vraisemblance L est la probabilité d'observer des données $\boldsymbol{y}$ en fonction du vecteur de paramètres $\boldsymbol{\theta}$ du modèle. La méthode du maximum de vraisemblance consiste à rechercher la valeur du paramètre $\boldsymbol{\theta}$ qui maximise la probabilité d'obtenir les données observées. En d'autres termes, cette méthode cherche la valeur du vecteur $\hat{\boldsymbol{\theta}}$ dans l'espace des paramètres Ω qui maximise la vraisemblance L. La fonction de vraisemblance L a la même expression que la densité de probabilité f, mais les observations θ_i y sont fixes et les probabilités π_i varient comme suit : $L(\boldsymbol{\theta};\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\theta})$. Comme la fonction logarithme est monotone, cette recherche équivaut à maximiser le logarithme l de L.

$$l(\hat{\boldsymbol{\theta}};\boldsymbol{y}) \geqslant l(\boldsymbol{\theta};\boldsymbol{y}) \qquad \forall \boldsymbol{\theta} \in \Omega \tag{6.15}$$

L'estimateur $\hat{\boldsymbol{\theta}}$ est généralement obtenu en différentiant l en fonction des θ_j et en résolvant le système d'équations (une équation par paramètre θ_j) :

$$\frac{\partial l}{\partial \theta_j} = 0 \qquad \forall j \in 1,\ldots,p \tag{6.16}$$

Si $\hat{\boldsymbol{\theta}}$ correspond à un maximum de l, la matrice des dérivées secondes :

$$\frac{\partial^2 l}{\partial \theta_j \partial \theta_j} \tag{6.17}$$

doit être définie négative pour $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Notons finalement la propriété d'invariance : si $g(\boldsymbol{\theta})$ est une fonction des paramètres $\boldsymbol{\theta}$, alors l'estimateur du maximum de vraisemblance de $g(\boldsymbol{\theta})$ est $g(\hat{\boldsymbol{\theta}})$. Par conséquent, nous pouvons utiliser une fonction des paramètres qui facilite l'estimation du maximum de vraisemblance, puis calculer l'estimation du maximum de vraisemblance pour les paramètres requis.

Régression logistique

La fonction de log-vraisemblance d'un modèle logistique s'écrit :

$$l(y_i,\ldots,y_n;\pi_1,\ldots,\pi_n) = \sum_{i=1}^{n} \left(y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} \right) - n_i \ln\left(1 + \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right) + \ln C_{y_i}^{n_i} \right) \tag{6.18}$$

La méthode du maximum de vraisemblance consiste à dériver l en fonction de β :

$$\begin{aligned} U_0 &= \frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - n_i \pi_i) \\ U_1 &= \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} x_{1i}(y_i - n_i \pi_i) \\ &\cdots \\ U_q &= \frac{\partial l}{\partial \beta_q} = \sum_{i=1}^{n} x_{qi}(y_i - n_i \pi_i) \end{aligned} \tag{6.19}$$

et à calculer la matrice d'information $\mathcal{J}$ $(1 \leqslant i \leqslant n)$:

$$\mathcal{J} = \begin{bmatrix} \sum n_i \pi_i (1-\pi) & \sum n_i x_{1i} \pi_i (1-\pi) & \dots & \sum n_i x_{qi} \pi_i (1-\pi) \\ \sum n_i x_{1i} \pi_i (1-\pi) & \sum n_i x_{1i}^2 \pi_i (1-\pi) & \dots & \sum n_i x_{1i} x_{qi} \pi_i (1-\pi) \\ \vdots & \vdots & \ddots & \vdots \\ \sum n_i x_{qi} \pi_i (1-\pi) & \sum n_i x_{qi} x_{1i} \pi_i (1-\pi) & \dots & \sum n_i x_{qi}^2 \pi_i (1-\pi) \end{bmatrix} \tag{6.20}$$

$\mathcal{J}$ peut s'écrire de manière compacte :

$$\mathcal{J} = \boldsymbol{X}^T W \boldsymbol{X} \tag{6.21}$$

$$W_{ii} = n_i \pi_i (1-\pi_i) \qquad W \text{ est diagonale} \tag{6.22}$$

L'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\beta}}$ est obtenu en résolvant l'équation itérative :

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + \left[\boldsymbol{\mathcal{J}}^{(m-1)}\right]^{-1} \boldsymbol{U}^{(m-1)} \tag{6.23}$$

où $\boldsymbol{b}$ est le vecteur estimant $\boldsymbol{\beta}$ et (m) indique la m$^{\text{e}}$ estimation.

0. Le calcul commence avec $b_0 = b_1 = \dots b_q = 0$, ce qui correspond à $\pi_i = 0.5 \; \forall i$,
1. Calcul de π_i à partir de $\boldsymbol{b}^{(m)}$ (eq. 6.13),
2. Calculs de $\boldsymbol{U}$ et $\mathcal{J}$ (eq. 6.19 et 6.20),
3. Si possible, calcul de l'inverse de $\mathcal{J}$.
   Sinon (cas où $\mathcal{J}$ est singulière), le calcul s'arrête sans converger.
4. Mise à jour de $\boldsymbol{b}$: $\boldsymbol{b}^{(m+1)}$
5. Si $\left| b_j^{(m+1)} - b_j^{(m)} \right| < \varepsilon$ pour $\forall 0 \leqslant j \leqslant q$, la convergence est atteinte.
   Sinon retour au point 1.

Samβada utilise cet algorithme pour estimer $\boldsymbol{\beta}$. Comme la méthode de détection de la sélection est basée sur les individus, chaque échantillon est considéré séparément ; il y a donc n points de mesure pour lesquels $n_i = 1$ et $y_i \in \{0, 1\}$. Cette approche permet de modéliser la probabilité qu'un individu porte l'allèle considéré en fonction de la composition environnementale (une ou plusieurs variables) de son habitat.

Dans la suite de la section, j'utiliserai la notation suivante :

n	Nombre d'échantillons
k	Nombre de variables environmentales
m	Nombre de marqueurs génétiques binaires

Pour chaque échantillon i dans un modèles comprenant q paramètres :

π_i	probabilité de présence du marqueur
$\boldsymbol{x}_i$	vecteur des variables environmentales
$\boldsymbol{\beta}$	vecteur des paramètres à $q+1$ coefficients

Évaluation de la significativité des modèles

Une fois qu'un modèle logistique est ajusté, il faut déterminer s'il représente une association significative entre le marqueur génétique et les variables environnementales. Samβada calcule deux scores pour évaluer la significativité des modèles. Une association est désignée comme significative si les deux tests statistiques rejettent l'hypothèse nulle.

Test du rapport de vraisemblance Le rapport des vraisemblances compare le modèle considéré avec un modèle de référence plus simple impliquant moins de variables explicatives. Dans le cas univarié, le modèle de référence est le modèle constant où la probabilité d'occurrence est égale à la fréquence moyenne du marqueur dans le jeu de données. L'hypothèse nulle (H_0) de ce test est que le modèle considéré n'explique pas mieux la distribution du marqueur que le modèle de référence.

$$G = 2\ln\left(\frac{L}{L_0}\right) = 2\cdot(l - l_0) \tag{6.24}$$

où :

L	vraisemblance du modèle courant	$l = \ln(L)$
L_0	vraisemblance du modèle de référence	$l_0 = \ln(L_0)$

Si l'hypothèse nulle est vraie, G suit une loi du χ^2 à h degrés de liberté. h est le nombre de paramètres ajoutés par rapport au modèle de référence, $h = q - q_0$. Si le modèle de référence n'a aucun paramètre (« modèle nul »), la log-vraisemblance s'écrit

$$l_0 = y\ln y + (n-y)\ln(n-y) - n\ln n \tag{6.25}$$

où le marqueur est présent y fois parmis n échantillons (Dobson et Barnett, 2008, p. 137).

Test de Wald La statistique de Wald détermine si les paramètres ajoutés au modèle courant par rapport au modèle de référence sont tous différents de 0.

$$\mathcal{W} = \left(\boldsymbol{\beta}^\star\right)^T \left(\left(\mathcal{J}^{-1}\right)^\star\right)^{-1} \boldsymbol{\beta}^\star \tag{6.26}$$

$\mathcal{J}$	matrice d'information (eq. 6.20)
$\mathcal{J}^{-1}$	matrice de variance-covariance
$\left(\mathcal{J}^{-1}\right)^{\star}$	$h \times h$ sous-matrice de $\mathcal{J}^{-1}$ concernant les paramètres ajoutés
$\boldsymbol{\beta}^{\star}$	sous-vecteur de $\boldsymbol{\beta}$, h paramètres ajoutés

Sous cette forme, le score de Wald est également comparé à une statistique du χ^2 à h degrés de liberté (Dobson et Barnett, 2008, p. 77).

Les scores G et de Wald sont généralement du même ordre de grandeur car ils convergent asymptotiquement (Engle, 1983). Il convient également de souligner que les scores G et de Wald ne mesurent pas la robustesse d'association, mais servent à tester la significativité de la corrélation.

Tests de significativité pour comparaisons multiples

Le niveau α d'un test fournit une borne supérieure à la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie.

$$\mathbb{P}(\text{faux rejet}) \leqslant \alpha \tag{6.27}$$

Ce niveau est généralement choisi pour réaliser un seul test.

Si de nombreux tests sont effectués simultanément, par exemple lors d'un balayage du génome, le niveau du test doit être modifié pour tenir compte de ces comparaisons multiples (Morgenthaler, 2007). En effet, supposons que l'on fasse l tests pour lesquels l'hypothèse nulle est vraie.

$$\begin{aligned}
\mathbb{P}(\text{aucun faux rejet}) &= \mathbb{P}\left(\cap_{i=1}^{l}(\text{le test } i \text{ ne rejette pas l'hypothèse nulle})\right) && (6.28)\\
&= 1-\mathbb{P}\left(\cup_{i=1}^{l}(\text{le test } i \text{ rejette l'hypothèse nulle})\right) && (6.29)\\
&\geqslant 1-l\alpha && (6.30)
\end{aligned}$$

Cette *inégalité de Bonferroni* (Bonferroni, 1936) permet de borner la probabilité de rejeter à tort l'hypothèse nulle :

$$\mathbb{P}(\text{au moins un faux rejet faux}) \leqslant l\alpha \tag{6.31}$$

où α est le niveau de chaque test individuel.

Plusieurs approches permettent de limiter le nombre de faux rejets lorsque de nombreux tests sont effectués simultanément.

Correction de Bonferroni La correction de Bonferroni contrôle la probabilité de faire au moins une fausse découverte, la *Family-Wise Error Rate* (*FWER,* Benjamini et Hochberg, 1995).

Elle consiste à modifier le niveau de chaque test individuel :

$$\alpha' = \alpha / l \tag{6.32}$$

Ainsi l'inégalité de Bonferroni 6.31 est bornée par α, le niveau souhaité pour l'ensemble des tests.

Remarquons ici que lorsque le nombre de comparaisons est élevé, la correction de Bonferroni conduit à des tests de significativité très conservateurs.

Taux de fausses découvertes Le contrôle de la probabilité de faire au moins une fausse découverte induit des faux négatifs parmi les modèles testés. Une procédure conservatrice est nécessaire en médecine pour tester l'efficacité de plusieurs nouveaux traitements par rapport à un produit couramment utilisé. En revanche, si le cadre de l'expérience permet d'accepter quelques faux positifs, la puissance du test peut être améliorée en choisissant une procédure de validation plus libérale. Les analyses génétiques par balayage du génome se situent dans ce contexte.

Un test de significativité peut déboucher sur quatre états possibles : l'hypothèse nulle peut être vraie ou fausse, et le test peut l'avoir rejetée ou non. Lors d'une l'analyse simultanée de l caractéristiques, le nombre de tests individuels appartenant à chaque catégorie est résumé par la table 6.1. Les corrections de type *family-wise* contrôlent que $\mathbb{P}(F \geqslant 1) \leqslant \alpha$. Elles ne

	Résultat déclaré significatif	Résultat non-significatif	Total
Hyp. nulle vraie	F	$l_0 - F$	l_0
Hyp. alternative vraie	V	$l_1 - V$	l_1
Total	S	$l - S$	l

Table 6.1 – Résultats possibles lors du test simultané de l caractéristiques. Les expériences sont classées en quatre groupes, l'hypothèse nulle peut être vraie ou fausse et le test peut l'avoir rejetée ou non.

renseignent pas sur le nombre de fois où l'hypothèse nulle était vraie parmi tous les tests.

Benjamini et Hochberg (1995) proposent une procédure pour contrôler le nombre de fois où l'hypothèse nulle a été rejetée à tort. Ils définissent le *taux de faux positifs* (*false discovery rate, FDR*) comme :

$$Q_e = \mathbb{E}\left[\frac{F}{F+V}\right] = \mathbb{E}\left[\frac{F}{S}\right] \tag{6.33}$$

Leur approche est équivalente au FWER si toutes les hypothèses nulles sont vraies, et a l'avantage d'être plus puissante que le FWER si certaines hypothèses nulles sont fausses.

La comparaison simultanée de l caractéristiques selon Benjamini et Hochberg se déroule ainsi :

1. Soient $H_1, \ldots, H_l$, les l hypothèses nulles et $P_1, \ldots, P_l$, les p-valeurs associées.

2. On note $P_{(1)} \leqslant P_{(2)} \leqslant \cdots \leqslant P_{(l)}$ les p-valeurs ordonnées et $H_{(1)}, H_{(2)}, \ldots, H_{(l)}$ les hypothèses nulles correspondantes.
3. On définit la procédure de comparaison multiple de type Bonferroni suivante :

   $$\text{Soit } k \text{ le plus grand } i \text{ tel que } P_{(i)} \leqslant \frac{i}{l} q^\star$$

   Alors on rejette tous les $H_{(i)}$ pour $i = 1, 2, \ldots k$.
4. Cette procédure contrôle le FDR au niveau $q^\star$ si les statistiques de tests sont indépendantes et pour n'importe quelle configuration d'hypothèses nulles.

On remarque que le seuil de significativité est augmenté chaque fois d'une hypothèse est rejetée. Le modèle ayant la plus petite hypothèse nulle doit cependant être significatif en appliquant la correction de Bonferroni.

Storey et Tibshirani (2003) présentent une procédure alternative pour calculer le taux de faux positifs $\mathbb{E}\left[\frac{F}{S}\right]$. Leur méthode permet d'associer une mesure de significativité à chaque modèle testé. Les p-valeurs sont à nouveau triées par ordre croissant. Si un modèle est considéré comme significatif, tous les modèles présentant une p-valeur plus petite ou égale seront aussi considérés comme significatifs. Storey et Tibshirani définissent la q-valeur d'un modèle comme la proportion de faux positifs induite en désignant ce modèle comme significatif. Par conséquent, en calculant la q-valeur de chaque modèle puis en fixant un seuil de significativité α, on obtient un ensemble de modèles significatifs dont une proportion α sont des faux positifs.

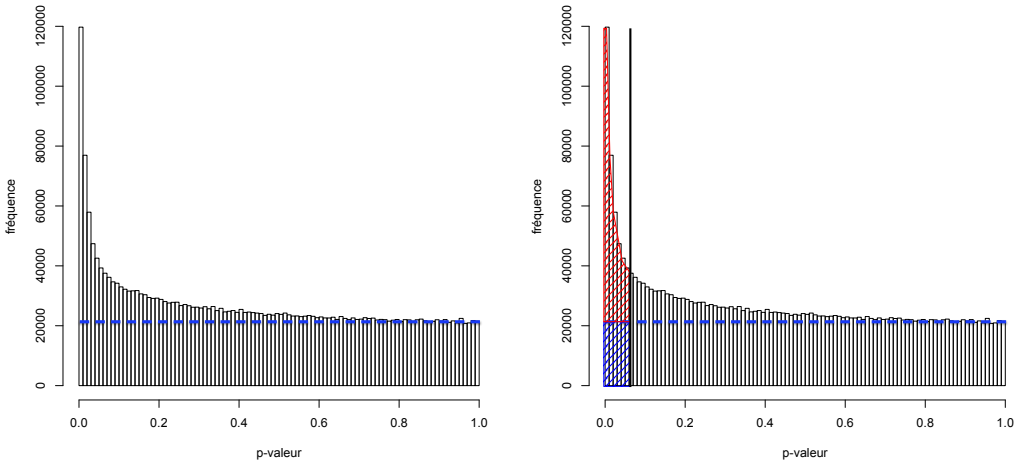
La procédure de Storey et Tibshirani se base sur une estimation empirique du taux de faux positifs. Un modèle pour lequel l'hypothèse nulle est vraie peut avoir une p-valeur uniformément distribuée entre 0 et 1. En revanche, un modèle pour lequel l'hypothèse nulle est fausse aura une p-valeur proche de 0. Le nombre de modèles pour lesquels l'hypothèse nulle est fausse (ou pour lesquels elle est vraie) peut donc être estimé à partir de l'histogramme des p-valeurs des l modèles testés (voir fig. 6.1). Le nombre de modèles pour lesquels l'hypothèse nulle est vraie (l_0) peut être estimé en comptant les tests ayant une p-valeur plus grande qu'un seuil λ (par ex. $\lambda = 0.5$) et en extrapolant ce nombre à l'ensemble de l'intervalle $[0;1]$. Ceci permet d'estimer le nombre de tests pour lesquels l'hypothèse nulle est vraie alors qu'ils possèdent une p-valeur plus petite que le seuil de significativité fixé. Ce sont les F faux positifs de la table 6.1 (p. 78).

Storey et Tibshirani proposent de calculer la q-valeur de chaque test de la manière suivante :

1. Soient $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(l)}$, les p-valeurs ordonnées des l tests.
2. Pour une suite de paramètres λ, par exemple $\lambda = \{0,01; 0,02; \ldots; 0,95\}$ calculer la proportion d'hypothèses nulles vraies :

   $$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1-\lambda)}$$

3. Soit $\hat{f}$ la spline cubique naturelle de $\hat{\pi}_0(\lambda)$ sur λ à 3 degrés de liberté.



(a) Distribution des p-valeurs

(b) Taux de fausses découvertes

Figure 6.1 – Histogramme des p-valeurs de l tests simultanés. *a)* Les tests pour lesquels l'hypothèse nulle est vraie ont une p-valeur uniformément répartie entre 0 et 1, alors que les tests pour lesquelles elle est fausse ont une p-valeur proche de 0. C'est pourquoi l'histogramme présente un pic pour les p-valeurs proches de 0, puis décroît jusqu'à devenir uniforme pour des p-valeurs proches de 1. Le trait discontinu bleu estime la distribution des p-valeurs pour les tests vérifiant l'hypothèse nulle, il est ajusté à partir de la fréquence des tests ayant des p-valeurs proches de 1. *b)* Lors du choix d'un seuil de significativité (barre verticale), la proportion de faux positifs parmi les modèles significatifs est estimée par le ratio entre la surface bleue (fausses découvertes) et la surface colorée totale (les vraies découvertes sont en rouge).

4. Soit l'estimateur de π_0 :
$$\hat{\pi}_0 = \hat{f}(1)$$
5. Calculer
$$\hat{q}(p_{(m)}) = \min_{t \geqslant p_{(m)}} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_j \leqslant t\}} = \hat{\pi}_0 \cdot p_{(m)}$$
6. Pour $i = m-1, m-2, \ldots, 1$, calculer
$$\hat{q}(p_{(i)}) = \min_{t \geqslant p_{(i)}} \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_j \leqslant t\}} = \min\left(\frac{\hat{\pi}_0 \cdot m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)})\right)$$
7. La q-valeur estimée pour le i^{e} modèle le plus significatif est $\hat{q}(p_{(i)})$.

Cette procédure permet d'obtenir une estimation facilement interprétable du nombre de faux positifs, mais elle requiert de connaître la distribution des p-valeurs pour les l tests.

Remarque : Samβada utilise la correction de Bonferroni pour les comparaisons multiples. Les FDR utilisés pour l'analyse des résultats ont été calculés dans R (voir sec. 7.1.3). En cas d'intérêt des utilisateurs, Samβada intégrera une procédure d'estimation du taux de faux positifs dans une prochaine version.

Autres critères pour la sélection de modèles

Outre les scores déjà présentés, deux statistiques sont utilisées pour comparer les modèles. Ces statistiques sont basées sur la log-vraisemblance et intègrent une pénalité proportionnelle au nombre de paramètres inclus dans les modèles. Le premier est le critère d'information d'Akaike (*Akaike information criterion*, AIC).

$$\mathrm{AIC} = -2l + 2p \tag{6.34}$$

où l est la log-vraisemblance du modèle et p est le nombre de paramètres.

La deuxième statistique est le critère d'information bayésien (*Bayesian information criterion*, BIC) où la pénalité dépend aussi du nombre d'observations.

$$\mathrm{BIC} = -2l + 2p \cdot \ln n \tag{6.35}$$

Sélection de modèles multivariés

Pour évaluer la significativité des modèles multivariés il est nécessaire d'adapter la méthode utilisée dans le cas univarié. S'il y a par exemple cinq variables environnementales $\{E_1, \ldots, E_5\}$ et qu'on considère des modèles multivariés, il s'agit de déterminer avec quel modèle M$\{E_1, E_3\}$ doit être comparé.

Une solution possible serait d'évaluer la significativité par rapport au modèle constant M_0 qui ne contient pas de variable environnementale. Dans ce cas, le score G indiquerait si le couple de variable $\{E_1, E_3\}$ fournit une meilleure prédiction de la présence du marqueur que le modèle constant, mais ne renseignerait pas sur la pertinence d'utiliser deux variables plutôt qu'une. De la même manière, le score de Wald mesurerait si les coefficients de régression correspondants aux deux variables peuvent être considérés comme nuls, et le rejet de l'hypothèse nulle montrerait qu'au moins un coefficient est différent de 0, ceci sans fournir d'indication sur les coefficients pris individuellement. Par conséquent, cette méthode ne permet pas de déterminer si le modèle M$\{E_1, E_3\}$ fournit une meilleure prédiction de la présence du marqueur que les modèles de dimension 1.

C'est pourquoi j'ai adopté une approche basée sur les parents. Pour rappel, les *parents* d'un modèle M à q variables sont les modèles de dimension $q-1$ (comprenant $q-1$ variables) se rapportant au même marqueur génétique et obtenus en écartant une variable de M (cf sec. 5.3.2 p. 67). En reprenant l'exemple ci-dessus, le modèle M$\{E_1, E_3\}$ a pour parents M$\{E_1\}$ et M$\{E_3\}$ (cf fig. 5.3 p. 66). La significativité d'un modèle multivarié est d'abord évaluée au moyen d'un rapport de vraisemblance similaire au score G pour les modèles univariés (cf éq. 6.24). Le test sélectionne le parent de M qui présente la log-vraisemblance la plus élevée, puis calcule

la différence entre la log-vraisemblance de M et celle de son parent.

$$G = 2\ln\left(\frac{L}{L_p}\right) = 2 \cdot (l - l_p) \tag{6.36}$$

où :
- L vraisemblance du modèle courant M — $l = \ln(L)$
- L_p vraisemblance du parent ayant la log-vraisemblance la plus élevée — $l_p = \ln(L_p)$

Cette différence est donc la plus petite qu'on puisse obtenir en comparant M avec un de ses parents. Si H_0 est vraie, G suit une loi de χ^2 à un degré de liberté car M comporte un paramètre de plus que ses parents. De cette manière, si M est significatif pour le test G présenté, il est aussi significatif en prenant un autre parent comme modèle de référence. La significativité de M est ensuite évaluée avec un test de Wald visant à déterminer si chaque coefficient de régression est non-nul (cf 6.37). Le score de Wald est donc calculé individuellement pour chaque coefficient j (parmi q coefficients). Dans ce cas le sous-vecteur $\boldsymbol{\beta}^\star$ et la sous-matrice $\mathcal{J}^{-1}$ comptent chacun un élément et le test de Wald s'écrit [1] :

$$\mathcal{W}_j = \beta_j \left(\left(\mathcal{J}^{-1}\right)_{jj}\right)^{-1} \beta_j = \frac{\beta_j^2}{\left(\mathcal{J}^{-1}\right)_{jj}} \quad 1 \leqslant j \leqslant q \tag{6.37}$$

où :
- $\mathcal{J}^{-1}$ matrice de variance-covariance
- β_j élément j du vecteur de paramètres $\boldsymbol{\beta}$

Si l'hypothèse nulle est vraie, chaque score $\mathcal{W}_j$ suit une distribution du χ^2 à un degré de liberté. Il suffit alors de tester si le plus petit score $\mathcal{W}_j$ est rejeté par le test, ce qui indique que chaque coefficient β_j est différent de zéro.

Samβada évalue la significativité des modèles multivariés avec ces deux tests et trie les modèles sélectionnés selon leur score de Wald.

6.1.2 Mesure de l'autocorrélation

Selon la « Première Loi de la Géographie » énoncée par Waldo. R Tobler (1970) *« Everything is related to everything else, but near things are more related to each other »*. Des événements spatialement proches ont plus de chances de se ressembler que s'ils avaient été éloignés. Cependant, les statistiques classiques requièrent que les échantillons étudiés soient indépendants (Dobson et Barnett, 2008, p. 51). La position des objets dans l'espace n'est pas neutre d'un point de vue statistique, et la valeur de leurs attributs dépend de cette localisation géographique (dépendance spatiale). Cette dépendance spatiale ou similitude (dissemblance) entre des événements due à leur proximité géographique peut être quantifiée pour déterminer si les valeurs observées dépendent de la position géographique. Cette mesure de l'autocorrélation

1. Remarque : le modèle compte q variables, mais il y a q+1 coefficients de régression car β_0 est le terme constant qui n'est pas pris en compte dans ce test. De même la matrice de variance-covariance $\mathcal{J}^{-1}$ est de dimension $(j+1) \times (j+1)$ mais $\left(\mathcal{J}^{-1}\right)_{00}$ n'est pas utilisé.

spatiale peut permettre d'éviter de tirer de fausses conclusions.

Deux indices d'autocorrélation spatiale sont implémentés dans Samβada : le I de Moran global (Moran, 1950) et l'indice local d'association spatiale (LISA) proposé par Anselin (1995). Ces mesures ont pour but de caractériser la distribution spatiale des loci potentiellement soumis à la sélection, notamment de localiser les régions où la présence d'un allèle chez un individu est corrélée à la fréquence de cet allèle chez ses voisins. Si un marqueur est significativement associé à un ou plusieurs facteurs environnementaux alors qu'il ne présente pas d'autocorrélation spatiale, cette association n'est vraisemblablement pas due à un effet démographique. Ce marqueur est alors un candidat à la sélection et il convient d'analyser sa fonction biologique en détail pour déterminer s'il a une valeur adaptative. La mesure de l'autocorrélation spatiale est une aide à l'interprétation des résultats de Samβada.

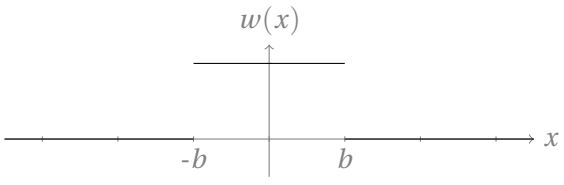
Pondération

La procédure générale de calcul de l'autocorrélation spatiale commence par la définition d'un voisinage (ou schéma de pondération) à considérer autour des points de mesure. Il s'agit d'une hypothèse sur le rayon d'influence des individus statistiques considérés. Le niveau d'association correspond à la corrélation entre la valeur d'un point et la moyenne pondérée des valeurs des points inclus dans le schéma de pondération défini.

Si la pondération ne dépend que de la distance entre le point et son voisin, la fonction associée utilise un noyau fixe. Si la pondération est ajustée en fonction de la densité de points dans la zone, la fonction associée utilise un noyau variable. Samβada propose trois fonctions de pondérations à noyau fixe (FM, NG, NB) et une fonction à noyau variable (NPPV). Le poids du point j qui est un voisin de i est notée w_{ij}. La distance d est la *bande-passante* (*bandwidth*) qui détermine l'étendue du voisinage.

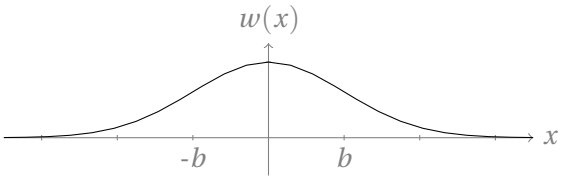
Fenêtre mobile (FM)

$$w_{ij} = \begin{cases} 1 & \text{si } d_{ij} < b \\ 0 & \text{sinon} \end{cases} \tag{6.38}$$



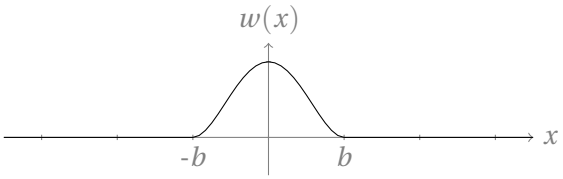
Noyau gaussien (NG)

$$w_{ij} = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \tag{6.39}$$



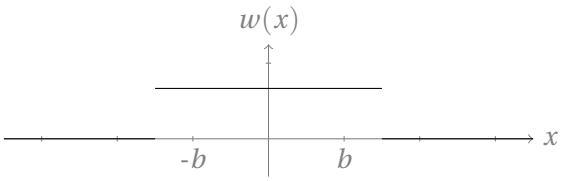
Noyau bicarré (NB)

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 & \text{si } d_{ij} < b \\ 0 & \text{sinon} \end{cases} \tag{6.40}$$



N plus proches voisins (NPPV)

$$w_{ij} = \begin{cases} \frac{1}{N} & \text{si } j \text{ est parmi les } N \text{ plus proches voisins de } i \\ 0 & \text{sinon} \end{cases} \tag{6.41}$$



En ce qui concerne l'implémentation de ces fonction de pondération dans Samβada, il faut noter que :

— Les pondérations des voisins d'un point sont normées pour que leur somme soit égale à 1. La somme de chaque ligne de la matrice de pondération W est alors égale à 1 ($\sum_{j \neq i} w_{ij} = 1$).

— S'il y a une valeur manquante pour une des variables dont Samβada doit mesurer l'autocorrélation, le logiciel recalcule la mise à l'échelle de chaque ligne de W pour cette variable.

— Dans le cas NPPV, si plusieurs points sont à la même distance de i que le N^{e} voisin, ils sont tous inclus dans le voisinage de i. La valeur de la pondération est ajustée en conséquence.

— Dans le cas NPPV, Samβada recalcule également la liste des voisins de chaque point si une des valeurs vient à manquer.

I de Moran global

L'indice I de Moran est une mesure de l'autocorrélation spatiale globale (Moran, 1950). Cela signifie qu'une mesure de dépendance spatiale entre les individus analysés est effectuée sur toute la zone géographique considérée.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{6.42}$$

avec

n	le nombre d'échantillons ;
S_0	la somme des poids $\left(S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}\right)$;
y_i, y_j	les valeurs aux emplacements i et j ;
$\bar{y}$	la moyenne ;
z_i, z_j	les écarts à la moyenne.

Le numérateur de l'éq. 6.42 est la somme pour tous les points i du produit entre l'écart (*deviation*) z_i et la somme pondérée des écarts des voisins $\sum_{j=1}^{n} w_{ij} z_j$. La valeur du I de Moran s'interprète ainsi :

$I = 0$	Pas d'autocorrélation spatiale, indépendance des échantillons, absence de dépendance spatiale ;
$I > 0$	Autocorrélation positive, similarité entre les voisins ;
$I < 0$	Autocorrélation négative, dissemblance entre les voisins.

Mesure de la significativité La quantification de la dépendance spatiale doit etre accompagnée d'une mesure de la significativité. En effet, il se peut que la mesure obtenue le soit par le hasard. Pour ce faire, on émet justement une hypothèse nulle supposant que la distribution des valeurs y_i parmi les points x_i est aléatoire. Comme la distribution de la statistique I n'a pas d'expression analytique, la méthode la plus courante pour déterminer si la valeur obtenue I_{obs} est compatible avec l'hypothèse nulle est de calculer une distribution empirique. La procédure est la suivante :

1. Effectuer R itérations ($1 \leqslant r \leqslant R$) :
   - Permuter les valeurs y_i aléatoirement entre les points x_i ;
   - Calculer I_{sim}^r.
2. pseudo p-valeur :
   - probabilité d'obtenir $I_{\text{sim}} \geqslant I_{\text{obs}} \geqslant \text{médiane}(I_{\text{sim}})$ (ou $I_{\text{sim}} \leqslant I_{\text{obs}} \leqslant \text{médiane}(I_{\text{sim}})$) par hasard ;
   - p-valeur $= (M+1)/(R+1)$, avec M évènements « I_{sim} plus extrême que I_{obs} ».

Remarque : Etant donné que $\mathbb{E}(I) = 1/(n-1)$, la médiane de la distribution empirique est généralement proche de 0. Pour le calcul de la significativité du I de Moran global, les

évènements où I_{sim} est plus extrême que I_{obs} peuvent être simplifiés en $I_{\text{sim}} \geqslant I_{\text{obs}}$ si $I_{\text{obs}} \geqslant 0$ et $I_{\text{sim}} \leqslant I_{\text{obs}}$ si $I_{\text{obs}} < 0$. De plus, il est très improbable de trouver $I_{\text{sim}} = I_{\text{obs}}$ et donc si la p-valeur obtenue est plus grande que 0,5, elle peut être corrigée en 1-p-valeur.

I de Moran local

Anselin (1995) a défini un *indicateur local d'association spatiale* (*Local Indicator of Spatial Association, LISA*) comme une statistique remplissant les deux conditions suivantes :

1. le LISA de chaque observation donne une indication de l'étendue de l'agglomération spatiale significative de valeurs similaires autour de cette observation ;
2. la somme des LISA pour toutes les observations est proportionnelle à un indicateur global d'association spatiale (I de Moran global).

La première propriété équivaut à tester la significativité statistique du LISA pour chaque observation.

Une version locale du I de Moran peut être obtenue à partir de l'éq. 6.42. Comme la somme de chaque ligne de W est égale à 1 $\left(\sum_{j_1}^{n} w_{ij} = 1\right)$, on a $S_0 = n$:

$$\begin{aligned} I &= \frac{\cancel{n}}{\cancel{S_0}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \\ &= \sum_{i=1}^{n} \frac{1}{\sum_{i=1}^{n} z_i^2} \cdot z_i \left(\sum_{j=1}^{n} w_{ij} z_j \right) \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \frac{n-1}{\sum_{i=1}^{n} z_i^2} \cdot z_i \left(\sum_{j=1}^{n} w_{ij} z_j \right) \end{aligned} \tag{6.43}$$

Ce qui mène à la définition suivante de l'indice de Moran local (Anselin, 1995) :

$$I_i = \frac{n-1}{\sum_{i=1}^{n} z_i^2} \quad \left(z_i \sum_{j=1}^{n} w_{ij} z_j \right) \tag{6.44}$$

$$\tag{6.45}$$

La somme des indices locaux est alors proportionnelle au I global :

$$I = \frac{1}{n-1} \sum_{i=1}^{n} I_i \tag{6.46}$$

Les indices locaux d'association spatiale permettent de caractériser le régime local d'autocorrélation spatiale. La taille des sous-régions considérées dépend du schéma de pondération. Leur valeur n'est pas forcément comprise entre 0 et 1.

Mesure de la significativité La mesure de la significativité des I de Moran locaux suit la même procédure que pour la version globale. Il faut cependant noter que :

1. Pour calculer la significativité de I_i, la valeur du point i reste fixe et les valeurs de ses voisins sont tirées parmi les $n-1$ valeurs disponibles.

2. Lors du calcul de l'autocorrélation locale de variables binaires, la pondération à fenêtre mobile et celle des plus proches voisins produisent fréquemment des indices simulés $I_{i,\text{sim}}$ égaux à la valeurs observée $I_{i,\text{obs}}$.

3. La distribution empirique des $I_{i,\text{sim}}$ n'est pas forcément centrée autour de 0. Il convient de compter séparément les cas $I_{i,\text{sim}} > I_{i,\text{obs}}$, $I_{i,\text{sim}} = I_{i,\text{obs}}$ et $I_{i,\text{sim}} < I_{i,\text{obs}}$ pour calculer la p-valeur en fonction de la position de $I_{i,\text{obs}}$ par rapport à la médiane de la distribution.

I de Moran bivarié

L'association spatiale de deux variables peut être mesurée comme la corrélation entre la valeur de l'une au point i et la moyenne pondérée de la valeur de l'autre pour les points voisins de i. En notant z et v les écarts des variables y et u par rapport à leur moyenne, l'indice de Moran bivarié s'écrit (Anselin et al., 2002) :

$$I_i(y,u) = \frac{n-1}{\sum_{i=1}^{n} z_i \cdot v_i} \quad \left(z_i \sum_{j=1}^{n} w_{ij} v_j \right) \tag{6.47}$$

$$I(y,u) = \frac{1}{n-1} \sum_{i=1}^{n} I_i \tag{6.48}$$

La significativité est mesurée selon le même principe que dans le cas univarié.

6.2 Méthodes corrélatives démographiques

Cette section présente deux méthodes corrélatives qui tiennent compte de la structure de population, et qui seront employées dans l'étude des bovins ougandais au chapitre suivant.

6.2.1 `BayEnv`

`BayEnv` implémente une approche de génomique environnementale basée sur les populations (Coop et al., 2010). Son but est de détecter des corrélations significatives entre les fréquences alléliques et des facteurs environnementaux. Or des populations géographiquement proches peuvent présenter des corrélations de leurs fréquences alléliques dues à leur histoire démographique commune. Comme ces populations proches peuvent également partager des conditions environnementales similaires, les corrélations peuvent mener à de fausses détections. `BayEnv` vise à détecter des signatures de la sélection naturelle tout en tenant compte non seulement de ces corrélations entre populations mais aussi des différences entre le nombre d'individus dans les populations. Le principe de `BayEnv` est d'utiliser un ensemble de loci de contrôle (sélectivement neutres) pour estimer une matrice de covariance des fréquences alléliques entre les populations (Ω). Cette matrice permet de reformuler l'hypothèse nulle lors

de la détection des signatures de sélection.

Coop et al. postulent qu'une population ancestrale s'est scindée en K sous-populations qui ont évolué séparément sous l'effet de la dérive génétique. L'histoire démographique n'est pas décrite en détail mais elle a pour effet que certaines populations présentent des corrélations entre leurs fréquences alléliques. Ces K populations sont échantillonnées pour L SNPs bialléliques. Les données pour un locus l se composent du nombre d'allèles de chaque type (1 et 2) échantillonnés dans chaque population. Les fréquences alléliques $(f_{1l},\ldots,f_{Kl})$ désignent celles de l'allèle 1. Le modèle suppose que la distribution conjointe des fréquences alléliques pour l'ensemble des populations suit une loi normale multidimensionnelle N dont la moyenne est la fréquence allélique ε_l dans la population ancestrale et dont la matrice de variance-covariance est la matrice de corrélation Ω multipliée par un terme propre au locus.

$$P(\theta_l|\Omega,\varepsilon_l) \sim N(\varepsilon_l,\varepsilon_l(1-\varepsilon_l)\Omega) \tag{6.49}$$

Le vecteur des fréquences alléliques θ_l est noté différemment des fréquences alléliques x_{kl} mesurées dans les populations car ces dernières sont comprises entre 0 et 1 alors que θ_l suit une distribution normale et peut légèrement déborder de cet intervalle. Il s'agit maintenant d'estimer la matrice Ω à partir des fréquences observées x_{kl} tout en tenant compte de l'incertitude concernant les paramètres inconnus ε_l et θ_l. Pour ce faire, `BayEnv` utilise une méthode de Monte-Carlo par chaînes de Markov (MCMC ; Coop et al., 2010).

Après avoir estimé un modèle nul pour la variation des fréquences alléliques entre les populations (la matrice Ω), `BayEnv` utilise ce modèle pour tester si les fréquences alléliques d'un SNP l sont corrélées significativement avec une variable environnementale X. Dans ce modèle alternatif, la fréquence allélique θ_l dépend cette fois-ci de la fréquence ancestrale ε_l et de la variable X. La déviation θ_l depuis la fréquence ancestrale est proportionnelle à X :

$$P(\theta_l|\Omega,\varepsilon_l) \sim N(\varepsilon_l+\beta X,\varepsilon_l(1-\varepsilon_l)\Omega) \tag{6.50}$$

Ce modèle doit être comparé au modèle nul (éq. 6.49) pour déterminer si la fréquence allélique est significativement corrélée à la variable environnementale. Cette comparaison est faite avec le facteur de Bayes qui mesure le soutien que les données apportent au modèle alternatif. Ce facteur est également calculé avec une méthode de Monte-Carlo par chaînes de Markov (Coop et al., 2010). Les modèles ayant les facteurs de Bayes les plus élevés sont les mieux étayés par les données. La significativité des modèles ne peut pas être calculée directement et nécessite de sélectionner les meilleurs modèles en fonction de la distribution empirique des facteurs de Bayes pour les loci neutres (voir sec. 7.1.4).

6.2.2 *Latent Factor Mixed Models* (`LFMM`)

Frichot et al. (2013) ont développé les *Latent Factor Mixed Models* (`LFMM`), une approche corrélative où la structure de population est introduite dans le modèle par l'intermédiaire de

variables non-observées. Ce modèle est basé sur les individus et permet de tester la significativité de la corrélation entre la présence des marqueurs et l'environnement, tout en estimant l'influence de ces variables cachées représentant la structure de population.

`LFMM` considère L loci bi-alléliques (p. ex. des SNPs) qui ont été échantillonnés pour n individus. Chaque SNP a un allèle ancestral et un allèle dérivé, et G_{il} est le nombre d'allèles dérivés pour le locus l de l'individu i. Pour des organismes diploïdes, G_{il} vaut 0, 1 ou 2. Les données comprennent également un vecteur de d variables environnementales X_i pour chaque individu. `LFMM` exprime le génotype G_{ij} avec un modèle linéaire mixte :

$$G_{il} = \mu_l + \beta_l^T X_i + U_i^T V_l + \epsilon_{il} \tag{6.51}$$

où :
- μ_l est un effet spécifique au locus l
- β_l^T est le vecteur des d coefficients de régression
- U_i est un vecteur de K facteurs latents propres à l'individu
- V_l est un vecteur de dimension K propre au locus
- ϵ_{il} est le résidu qui suit une distribution normale $N(0, \sigma^2)$

L'ensemble des termes $U_i^T V_l$ forment une matrice représentant la part de la variabilité génétique qui n'est pas expliquée par l'influence de l'environnement. Les paramètres sont ajustés dans un contexte bayésien, avec un échantillonneur de Gibbs (Dobson et Barnett, 2008). `LFMM` traite des modèles à une variable environnementale et évalue leur significativité avec le score $|z|$, soit le rapport entre la valeur centrée des coefficients β_l et leur écart-type. Si l'hypothèse nulle est vraie, c'est-à-dire si le SNP est neutre, le score $|z|$ suit une loi normale. La significativité de chaque modèle est ainsi estimée avec une p-valeur (Frichot et al., 2013). Le score $|z|$ est apparenté au score de Wald. Le nombre de facteurs latents K peut être estimé à partir d'une analyse de la structure de population. Les auteurs suggèrent d'utiliser la théorie de Tracy-Widom pour déterminer la valeur optimale de K (Patterson et al., 2006).

6.3 Génétique des populations

Cette section présente deux méthodes de génétique des populations utilisées dans cette thèse. La première permet d'analyser la structure d'une population et la seconde détecte les loci singuliers qui sont potentiellement soumis à la sélection naturelle. Elle seront également utilisées au chapitre suivant.

6.3.1 `Admixture`

`Admixture` permet d'analyser la structure d'une population (Alexander et al., 2009). Cette approche présume que les individus observés sont issus d'un nombre prédéfini de populations ancestrales qui se mélangent. `Admixture` permet d'estimer la fréquence des marqueurs génétiques dans les populations ancestrales et également la proportion du génome de chaque

individu qui provient de chacune de ces populations. Les données utilisées par `Admixture` sont des SNPs biallélliques, I individus sont génotypés pour J loci et il y a K populations ancestrales. La population k fournit une fraction q_{ik} du génome de l'individu i. Concernant les populations ancestrales, la fréquence de l'allèle 1 du SNP j dans la population k est f_{kj}. Les données disponibles sont les génotypes des individus, qui sont codés ainsi : g_{ij} compte le nombre d'allèles de type 1 présents au locus j de l'individu i ($g_{ij} = 0$, 1 ou 2). Les proportions q_{ik} et les fréquences f_{kj} sont inconnues et doivent être estimées à partir des données. `Admixture` présume que le génotype des individus est formé par une union aléatoire des gamètes à partir des populations ancestrales. La probabilité que l'individu i ait chaque génotype g_{ij} au locus j est donnée par un tirage binomial :

$$\begin{aligned} P(g_{ij} = 2) &= \left[\sum_k q_{ik} f_{kj}\right]^2 \\ P(g_{ij} = 1) &= 2\left[\sum_k q_{ik} f_{kj}\right]\left[\sum_k q_{ik}(1 - f_{kj})\right] \\ P(g_{ij} = 0) &= \left[\sum_k q_{ik}(1 - f_{kj})\right]^2 \end{aligned} \tag{6.52}$$

Dans cette équation, le produit $q_{ik} f_{kj}$ peut être considéré comme la « fraction » d'allèle 1 fournie par la population k. La somme $\sum_k q_{ik} f_{kj}$ est la probabilité que l'individu ait un allèle 1 à ce locus et elle est élevée au carré pour avoir la probabilité que l'individu ait le génotype 1/1 ($g_{ij} = 2$). Si tous les individus sont indépendants, la vraisemblance du modèle s'écrit :

$$L(Q,F) = \sum_i \sum_j \left\{ g_{ij} \ln\left[\sum_k q_{ik} f_{kj}\right] + (2 - g_{ij}) \ln\left[\sum_k q_{ik}(1 - f_{kj})\right] \right\} \tag{6.53}$$

Les matrices $Q = \{q_{ij}\}$ et $F = \{f_{kj}\}$ sont de dimensions $I \times K$ et $K \times J$. La première représente la proportion du génome de chaque individu issu de chaque population ancestrale et la deuxième représente la fréquence de chaque SNP dans ces K populations ancestrales. Ces matrices sont estimées à partir des génotypes des individus. `Admixture` repose sur les mêmes principes que `structure` (Pritchard et al., 2000) mais les calculs convergent beaucoup plus rapidement grâce à un algorithme de maximisation de la vraisemblance particulier qui est plus efficace que l'analyse bayésienne de `structure` (Alexander et al., 2009).

`Admixture` inclut une méthode de validation croisée pour déterminer la valeur optimale de K. Les génotypes sont répartis en Z groupes qui sont masqués tour à tour, les marqueurs masqués variant d'un individu à l'autre. A chaque tour, les coefficients d'appartenance et les fréquences alléliques dans les populations ancestrales sont calculés sur la base des génotypes visibles. Les valeurs des génotypes masqués sont ensuite prédites à partir des coefficients d'appartenance et des fréquences alléliques dans les populations ancestrales. L'erreur de validation croisée mesure la moyenne de la déviance entre le génotype prédit pour les valeurs cachées et le génotype réel. Cette analyse est répétée pour différentes valeurs de K. L'erreur de

validation croisée minimale désigne la structure de populations la plus probable (Alexander et al., 2013).

6.3.2 `Arlequin`

`Arlequin` est un logiciel polyvalent d'analyse de données en génétique des populations (Excoffier et Lischer, 2010). Nous décrivons ici uniquement les fonctions liées à la détection de loci sous sélection à partir de la diversité génétique observée entre populations (Excoffier et Lischer, 2011). L'approche utilisée par `Arlequin` est celle de Beaumont et Nichols (1996). Chaque locus est caractérisé par un indice de différenciation entre populations qui dépend du taux d'hétérozygotie (proportion d'individus hétérozygotes) à ce locus. La distribution de cet indice est simulée pour des loci sélectivement neutres sur de nombreuses générations, ce qui permet de détecter les loci présentant une valeur singulière de cet indice, et qui sont potentiellement soumis à la sélection. Cette approche utilise un modèle en îles formé de d populations de N individus. A chaque génération, quelques individus (et donc quelques allèles) migrent d'une île à l'autre selon un taux m. Chaque île reçoit en moyenne $N \cdot m$ migrants par génération [2]. `Arlequin` calcule la statistique Φ_{ST} pour les organismes étudiés en utilisant une analyse moléculaire de la variance (*Analysis of molecular variance* (AMOVA), Excoffier et al., 1992). Φ_{ST} est similaire à l'indice de fixation F_{ST} (Wright, 1949) et mesure la proportion de la diversité génétique qui est due à la différentiation entre populations (Holsinger et Weir, 2009). Le calcul de F_{ST} (ou Φ_{ST}) permet d'estimer le taux de migration (Slatkin, 1991) :

$$F_{ST} = \frac{1}{1 + \frac{4Nmd}{d-1}} \tag{6.54}$$

La simulation a lieu en deux étapes. La première commence avec d populations de $2N$ allèles (si les individus sont diploïdes). L'évolution est simulée en remontant dans le temps : à chaque génération, deux allèles peuvent fusionner. Ce processus s'appelle « coalescence » et modélise le fait que deux allèles peuvent provenir du même allèle de la génération précédente (c'est-à-dire qu'un individu peut avoir plusieurs descendants et leur transmettre le même allèle). Donc, en remontant dans le temps, certains allèles peuvent fusionner à chaque génération avec une probabilité plus grande s'ils appartiennent à la même population (car les migrations sont rares). Le processus se poursuit jusqu'à ce qu'il ne reste plus qu'un allèle. La simulation coalescente peut être représentée par un arbre dont certaines branches se rejoignent à chaque génération. La deuxième étape se déroule dans la direction réelle du temps. En partant du sommet de l'arbre (un allèle), des mutations sont introduites sur certaines branches, leur nombre dépendant du taux de mutation.Le modèle à nombre infini de sites (*infinite-site model*) suppose que chaque mutation se produit à un locus différent et permet de simuler facilement des SNPs bialléliques. Les $2N \cdot d$ allèles obtenus portent donc chacun une série de mutations qui découlent de leur généalogie dans l'arbre de coalescence. La structure en îles

2. `Arlequin` implémente aussi un modèle de populations en pierres de gué (*stepping stones*) où les populations échangent des migrants qu'avec leurs voisines. Ce modèle est une représentation plus réaliste de la propagation des nouvelles mutations d'une population à l'autre.

des populations simulées permet de calculer l'indice de fixation F_{ST} de chaque mutation (en relation avec son taux d'hétérozygotie). Ce processus coalescent est répété de nombreuses fois en faisant varier le taux de mutation. Ces simulations permettent d'estimer une distribution de F_{ST} pour des loci neutres en fonction de leur taux d'hétérozygotie. En comparant les données empiriques avec cette distribution, les loci qui présentent une valeur singulière de F_{ST} sont détectés comme potentiellement soumis à la sélection naturelle.

La principale différence entre `Arlequin` et d'autres logiciels de génétique des populations est que ce dernier permet de créer des structures hiérarchisées où les populations sont divisées en groupes. Les populations échangent plus de migrants à l'intérieur des groupes qu'entre les groupes (Excoffier et al., 2009). Ce modèle permet de simuler certains types de populations de manière plus réaliste.

La présentation d'`Arlequin` clôt ce chapitre consacré aux méthodes de détection des loci sous sélection et de caractérisation de la structure de population. Tous les éléments sont à présents réunis pour passer à l'analyse des données.

7 Identification de loci sous sélection chez *Bos taurus* et *Bos indicus*

Ce chapitre présente les résultats obtenus en Ouganda. Après un rappel des jeux de données disponibles, l'analyse commence avec la structure de population. Puis les loci soumis à la sélection naturelle sont recherchés avec quatre méthodes et les détections respectives sont comparées. L'analyse se poursuit avec l'étude de la distribution spatiale de trois SNPs. Le chapitre se clôt sur une courte étude de données simulées.

7.1 Ouganda

7.1.1 Données et méthodes utilisées

Les animaux échantillonnés en Ouganda ont été analysés selon deux groupes comme présenté dans la section 4.5.1. Le premier groupe de 813 individus a été génotypé avec une puce à moyenne densité (54k) SNPs et le deuxième groupe de 102 individus l'a été avec une puce à haute densidé (800k SNPs). Les données moléculaires ont été filtrées avec le logiciel `PLINK` pour obtenir les deux principaux jeux de données suivants [1] :

Données 54k soit 804 individus et 41'215 SNPs, fréquence allélique minimale (M. A. F.) : 1% ;

Données 800k soit 102 individus et 634'849 SNPs, M. A. F. : 5%.

De plus, comme nous voulions déterminer si une puce 54k pouvait être utilisée à la place d'une puce 800k dans une étude afin de génotyper plus d'échantillons avec le même budget, nous avons vérifié si les analyses basées sur la puce 54k fournissaient des résultats similaires à ceux fournis par la puce 800k. Pour ce faire, nous avons extrait les marqueurs présents sur la puce 54k du jeu de données 800k. Il s'avère que 5'264 marqueurs de la puce 54k ne sont pas inclus sur la puce 800k et que certains marqueurs présents sur les deux puces ont été écartés lors du filtrage, les marqueurs communs aux deux jeux de données sont donc moins nombreux que ceux du jeu 54k.

Données 800ksub soit 102 individus et 31'576 SNPs, M. A. F. : 5%.

1. En collaboration avec Pablo Orozco-terWengel (cf note p. 49).

Ces trois jeux de données sont répartis sur l'entier du génome et ont servi à analyser la structure de population. La recherche des signatures de sélection nécessite quelques précautions supplémentaires. En effet, les données utilisées doivent être restreintes aux autosomes (chromosomes 1 à 29 chez les bovins) car les chromosomes sexuels X et Y doivent être analysés à part. Cela provient du fait que tous les mammifères ont deux chromosomes homologues pour les autosomes, et que les femelles ont également deux chromosomes X alors que les mâles ont un X et un Y. Ainsi une population de mammifères comporte en moyenne moins de chromosomes X que d'autosomes (environ 0.75 pour 1) et encore moins de chromosomes Y (environ 0.25 pour 1 autosome). Cette différence de taille de population effective entre les chromosomes a deux effets sur l'évolution. D'une part, les maladies liées aux chromosomes X et Y sont dominantes chez les mâles car ils n'ont qu'un chromosome de chaque type ; la pression de sélection est donc plus intense sur ces chromosomes. D'autre part, comme les chromosomes sexuels sont moins nombreux que les autosomes, la variation aléatoire des fréquences alléliques lors de la reproduction est plus marquée ; la dérive génétique est donc plus rapide pour ces chromosomes. C'est pourquoi les chromosomes sexuels présentent des signatures de sélection plus nombreuses que les autosomes et ne peuvent donc pas être analysés en même temps.

En parallèle, certains SNPs sont référencé sur la puce comme appartenant au chromosome 0. Cela signifie que leur position sur le génome est inconnue. Ces SNPs peuvent également être écartés de l'analyse car ils n'apportent pas d'information utile.

Nous avons donc extrait les chromosomes 1 à 29 dans les trois jeux de données. Ce nouveau filtrage ne change pas les individus analysés et ne modifie pas non plus l'intersection des jeux 54k et 800k pour les loci autosomes. Nous avons donc recherché les signatures de sélection dans les données suivantes :

Données 54k soit 804 individus et 40'034 SNPs, fréquence allélique minimale (M. A. F.) : 1% ;

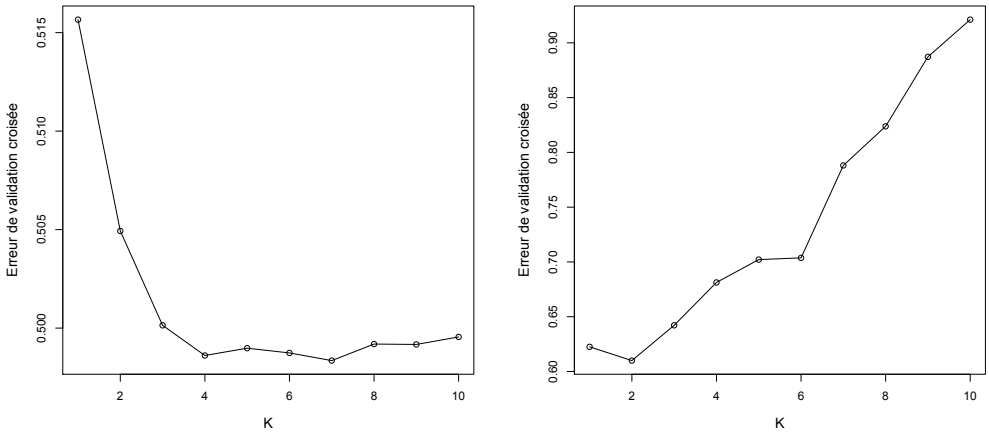
Données 800k soit 102 individus et 599'698 SNPs, M. A. F. : 5%.

Données 800ksub soit 102 individus et 30'679 SNPs, M. A. F. : 5%.

De manière à documenter les performances respectives des approches corrélatives récemment développées, j'ai recherché les marqueurs soumis à la sélection dans ces trois jeux de données avec trois méthodes de génomique environnementale, soit Samβada, `BayEnv` (Coop et al., 2010) et `LFMM` (Frichot et al., 2013), et une approche de génomique des populations, `Arlequin` (Excoffier et Lischer, 2010).

L'environnement est caractérisé par les 23 variables topo-climatiques décrites à la section 4.5.2.

Avant de passer à la détection des loci sous sélection, le paragraphe suivant traite de la structure de populations, information nécessaire au fonctionnement de `BayEnv` et `Arlequin` et utile à celui de `LFMM`.



(a) Données 54k

(b) Données 800k

Figure 7.1 – Validation croisée du classement effectué par `Admixture`. Le premier jeu de données contient 804 individus et 41'215 SNPs, alors que le second comprend 102 individus et 634'849 SNPs. L'axe des abscisses compte le nombre de populations considérées (K). L'axe des ordonnées indique l'erreur de validation croisée (voir sec. 6.3.1). La plus petite valeur de cette erreur correspond à la meilleure partition des individus.

7.1.2 Structure de populations

Nous avons utilisé `Admixture` (Alexander et al., 2009) pour analyser la structure des jeux de données 54k, 800k et 800ksub (sec. 7.1.1). En plus du filtrage présenté ci-dessus, nous avons écarté de l'analyse tous les SNPs ayant une valeur manquante dans le fichier `.map` qui décrit les loci utilisés et leur position sur le génome. En effet, `Admixture` requiert que ce dernier soit complet.

La figure 7.1 montre les résultats du test de validation croisée pour les données 54k et 800k. Les meilleures partitions sont $K = 4$ ou $K = 7$ pour les données 54k et $K = 2$ pour le jeux 800k. Avec les données 54k, le gain en précision du modèle est plus important lorsque l'on passe de trois à quatre populations que lorsque l'on passe de six à sept. De plus, le classement en quatre populations est plus parcimonieux que celui en sept populations, alors que ses performances sont presque aussi bonnes. Pour les données 800k, le classement en deux populations a le meilleur score de validation.

La distribution spatiale des quatre populations est présentée sur la figure 7.2. Les deux principaux clusters sont situés au sud-ouest et au nord-est du pays. Une des petites populations (cluster 1) se concentre principalement autour de Kampala [2], la capitale, avec quelques représentants à l'ouest, alors que les individus appartenant à l'autre population minoritaire se situent exclusivement dans l'ouest de l'Ouganda. La figure 7.3 présente quatre individus ayant

2. La capitale de l'Ouganda est située au sud-est du pays au bord du lac Victoria (voir carte générale p. 50).

un coefficient d'appartenance élevé aux clusters 2 ou 3. Les deux premières vaches (a et b) ont de longues cornes et une robe brune caractéristiques de la race ankole (*Bos taurus*). Cette population est principalement élevée dans le sud-ouest du pays. Les deux photos suivantes (c et d) présentent des individus ayant des cornes courtes et une bosse sur le garrot. Ce sont des zébus (*Bos indicus*) arrivés d'Inde à partir du VIII^{e} s. et qui vivent principalement dans le nord-est de l'Ouganda (Ajmone Marsan et al., 2010).

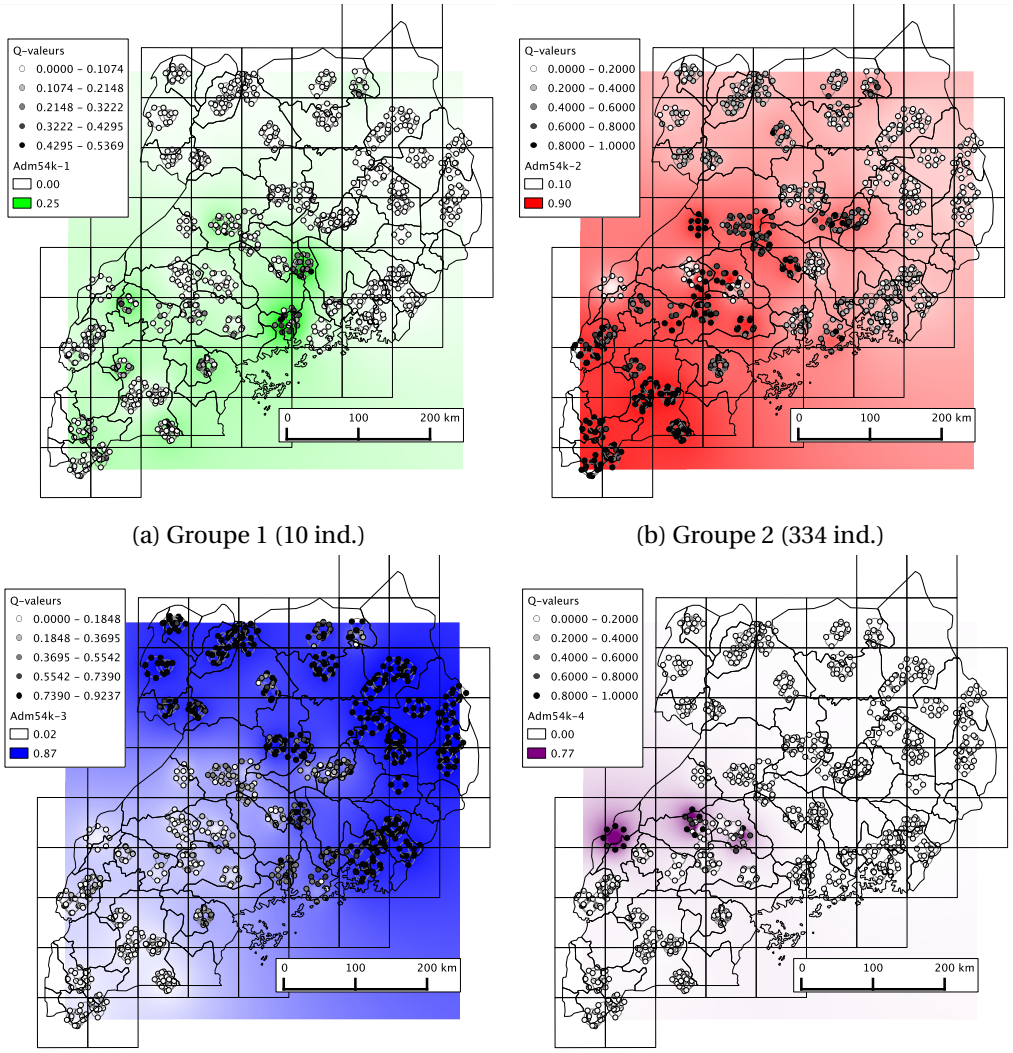
La structure de population est présentée sur la fig. 7.5. Les deux premiers graphiques depuis le haut montrent les partitions des 804 individus en quatre et sept populations et le troisième montre la partition des 102 individus en deux populations. Les individus sont attribués à la population de laquelle la plus grande fraction de leur patrimoine génétique est issue. A l'intérieur d'un cluster, ils sont classés par valeur croissante ou décroissante de cette fraction. L'ordre des échantillons n'est donc pas le même sur les graphiques 7.5a et 7.5b.

D'après la fig. 7.5c, la plupart des vaches ankoles appartiennent clairement à la première population. Puis la proportion du génome d'origine zébu croît linéairement jusqu'à devenir majoritaire. A l'inverse, presque tous les zébus ont une partie de leur patrimoine génétique d'origine ankole. La carte 7.4 présente la distribution spatiale de ces individus. Comme observé précédemment, les vaches ankoles peuplent le sud-ouest de l'Ouganda alors que les zébus sont principalement élevés dans le nord-est. L'analyse du sous-ensemble 800ksub par `Admixture` produit globalement la même partition. Seuls deux individus sont assignés différemment.

Ces analyses sont cohérentes avec les études préexistantes : les éleveurs ont des pratiques différentes suivant la race qu'ils possèdent. Les éleveurs d'ankoles ne les croisent que rarement avec les zébus car elles ont une valeur culturelle beaucoup plus importante. Des croisements ont cependant lieu avec des races étrangères pour tenter d'augmenter leur production laitière. En revanche, les éleveurs de zébus les hybrident couramment avec des ankoles afin d'augmenter leur résistance aux parasites. La pression de sélection due au trypanosome est suspectée de limiter la propagation des zébus vers le sud de l'Ouganda (Ajmone Marsan et al., 2010 ; Groeneveld et al., 2010).

La classification en quatre populations illustre également la séparation entre les zébus et les ankoles (7.5a). Les individus affiliés à la première population (tout à gauche de la figure) ont une large part de leur génome issue des clusters 2 et 3, alors que les individus de la quatrième population (tout à droite) présentent une hybridation moindre. A l'interface des populations ankole (n^{o} 2) et zébu (n^{o} 3), certains individus semblent être un croisement entre les quatre populations.

Dans la classification en sept populations (fig. 7.5b), les clusters 2, 3 et 7 sont distincts et comportent peu d'hybridation. Les ankoles (pop. 1) présentent un peu plus de gènes issus des autres populations que précédemment. Les individus appartenant à la population 4 possèdent une grande fraction de leur génome d'origine étrangère. La principale différence concerne les zébus qui se sont séparés en deux sous-populations. La majorité des zébus sont un croisement entre les populations 5 et 6. La carte 7.6 montre la distribution spatiale de ces



(a) Groupe 1 (10 ind.)

(b) Groupe 2 (334 ind.)

(c) Groupe 3 (438 ind.)

(d) Groupe 4 (22 ind.)

Figure 7.2 – Carte des quatre populations calculées par `Admixture` avec les données 54k et 804 individus. Ceux qui appartiennent à la même ferme ont été répartis sur un cercle entourant leur position réelle. Sur chaque carte, la couleur des points reflète la proportion du génome de l'individu qui est dérivée de la population ancestrale considérée. L'interpolation est calculée en fonction de l'inverse du carré de la distance et permet de situer les régions où ces populations sont les plus présentes afin d'améliorer la visualisation des classes d'appartenance. Les deux principaux clusters sont situés au sud-ouest et au nord-est de l'Ouganda.

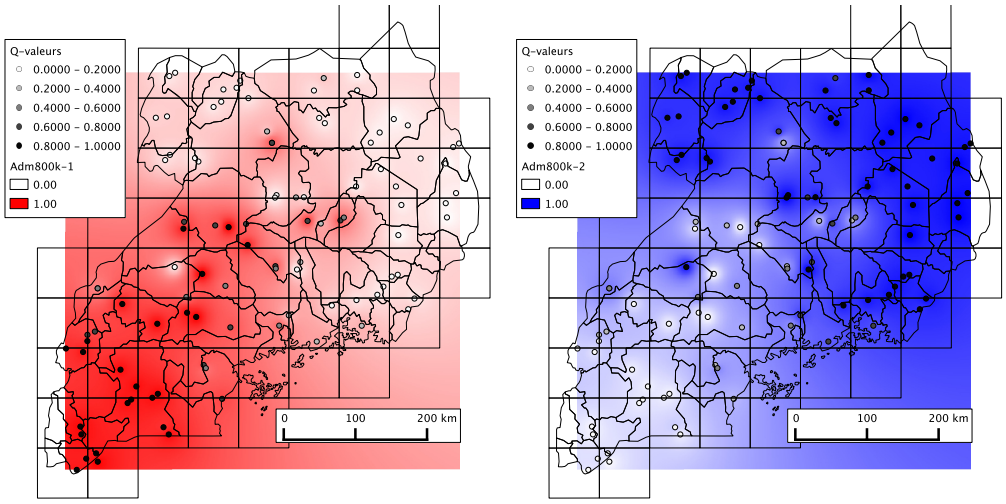
(a) Vache ankole près de Mpondwe, au nord du Lac Édouard (ouest du pays) $Q_{2,4} = 97,8\%$, $Q_{1,7} = 95,5\%$

(b) Vache ankole à l'ouest de Ntungamo (sud-ouest du pays) $Q_{2,4} = 97,0\%$, $Q_{1,7} = 92,3\%$

(c) Zébu près de Abim (nord-est du pays) $Q_{3,4} = 90,5\%$, $Q_{5,7} = 55,9\%$, $Q_{6,7} = 39,1\%$

(d) Zébu près de Arua (nord-ouest du pays) $Q_{3,4} = 84,5\%$, $Q_{6,7} = 99,9\%$

Figure 7.3 – Illustrations de vaches ankoles et de zébus. Les individus ont été choisis parmi les photographies NextGen disponibles en fonction des classifications en quatre et sept populations fournies par `Admixture` pour les données 54k. Le coefficient d'appartenance à la population 2 parmi 4 est noté $Q_{2,4}$.



(a) Groupe 1 (43 ind.) (b) Groupe 2 (59 ind.)

Figure 7.4 – Carte des deux populations calculées par `Admixture` avec les données 800k

deux populations de zébus. Le cluster 5 est principalement présent dans l'Est de l'Ouganda avec quelques représentants au nord de Kampala[3], alors que le cluster 6 se concentre au nord-ouest du pays avec quelques individus près de la capitale.

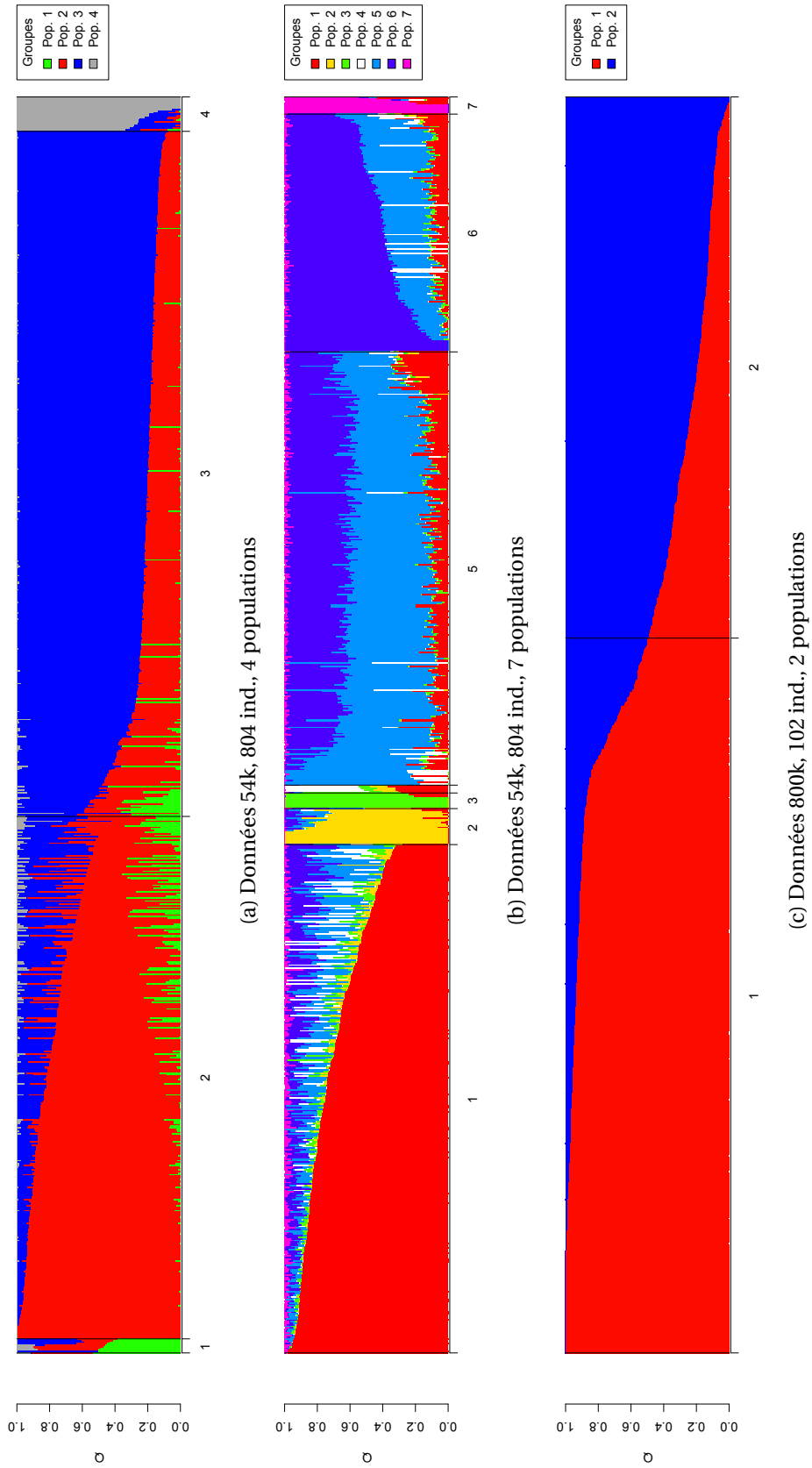
Les assignations des individus en populations pour les partitions $K = 4$ et $K = 7$ sont comparées sur la table 7.1. En passant de quatre à sept populations, le groupe 1_4 se sépare entre les clusters 4_7, 5_7 et 6_7. Le cluster 4_7 ne contient pas d'autres individus, alors que les clusters 5_7 et 6_7 contiennent en plus des zébus. Le cluster 1_4 semble donc plus apparenté aux zébus qu'aux ankoles. Il n'y a malheureusement pas de photos de ces individus. Le cluster 2_4 se retrouve principalement dans le groupe 1_7 avec quelques individus répartis dans les clusters 3_7 et 7_7 qui forment donc de petites populations d'ankoles. Le cluster 3_4 se divise entre les populations 5_7 et 6_7 comme observé précédemment. Quelques individus sont affiliés à la population ankole 1_7. La population 4_4 se retrouve entièrement dans le groupe 2_7. Les quelques photos disponibles présentent des vaches à la morphologie proche des ankoles.

Pop.		$K=7$						
		1	2	3	4	5	6	7
$K=4$	1	0	0	0	5	1	4	0
	2	313	0	10	0	0	0	11
	3	13	1	0	0	276	148	0
	4	0	22	0	0	0	0	0

Table 7.1 – Comparaison des classifications du jeu de données 54k en quatre et sept populations par `Admixture`. La situation représente 804 individus et 41'215 SNPs.

La table 7.2 compte le nombre d'individus assignés à chaque population pour cinq seuils

3. Voir carte générale p. 50.

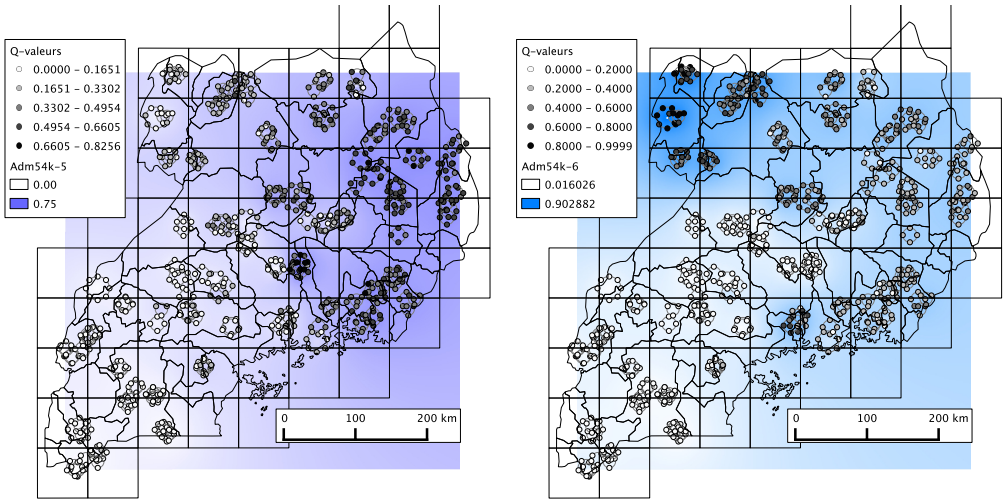


(a) Données 54k, 804 ind., 4 populations

(b) Données 54k, 804 ind., 7 populations

(c) Données 800k, 102 ind., 2 populations

Figure 7.5 – Structures de populations calculées avec `Admixture`. Les individus sont regroupés par populations, numérotées horizontalement. L'attribution est faite en fonction du plus haut coefficient d'appartenance Q_{max} de l'échantillon. A l'intérieur de chaque population, les individus sont classés par valeur croissante (ou décroissante) de Q_{max}. L'ordre des individus n'est donc pas le même pour les fig. a et b.



(a) Groupe 1 (43 ind.) (b) Groupe 2 (59 ind.)

Figure 7.6 – Carte des deux populations de zébus calculées par `Admixture` avec les données 54k pour $K = 7$.

d'appartenance Q. Peu d'individus appartiennent à une population avec un score supérieur à $Q = 0,85$. Les populations apparaissent vers $Q = 0,7$ et la plupart des individus sont rattachés à une population pour $Q = 0,5$. Dans le cas des sept populations basées sur les données 54k, certains individus sont très hybridés et ont un coefficient d'appartenance maximum de 0,27. Cela concerne un petit groupe de vaches ankoles et une part importante des zébus des clusters 5 et 6. Cette table montre que les croisements sont relativement fréquents entre les populations.

Il faut mentionner ici que lors de l'échantillonnage, la race de chaque animal a été répertoriée avec d'autres variables phénotypiques. Ces informations sont basées sur les déclarations des éleveurs et l'expertise de l'échantillonneur. La table 7.3 croise les races observées sur le terrain avec la structure de population estimée par `Admixture`. Cette comparaison confirme les résultats précédents en y apportant quelques nuances. En effet, les clusters 2_4, 1_7 et 1_2 correspondent bien à des vaches ankoles, mais certains individus répertoriés comme ankoles sont génétiquement plus proches des zébus. D'après leurs photos, ces animaux présentent parfois une morphologie hybride (par ex. robe brune et bosse sur le garrot) ou semblent parfois avoir été mal étiquetés. Les clusters 3_4, 5_7 et 6_7 sont effectivement composés de zébus. Quelques représentants de ces groupes ont été répertoriés comme ankoles.

La table 7.3 fournit également des indications sur les races minoritaires. Les Nganda semblent être un croisement entre ankoles et zébus, car elles sont partagées entre les clusters 1_4 et 3_4 ainsi qu'entre 1_7 et 6_7 (voir aussi Joshi et al., 1957). Les races Nkiga, Nsongora, Ntoro sont apparentées aux ankoles, alors que l'individu Ntuku est plus proche des zébus.

		Paliers Q					Total
		$Q \geqslant 0.99$	$Q \geqslant 0.85$	$Q \geqslant 0.70$	$Q \geqslant 0.50$	$Q \geqslant 0.30$	
Pop.	1	0	0	0	3	10	10
	2	9	149	243	321	334	334
	3	0	73	374	424	438	438
	4	8	10	20	22	22	22

(a) Données 54k, 4 populations.

		Paliers Q					Total
		$Q \geqslant 0.99$	$Q \geqslant 0.85$	$Q \geqslant 0.70$	$Q \geqslant 0.50$	$Q \geqslant 0.27$	
Pop.	1	0	92	193	275	326	326
	2	8	10	20	22	23	23
	3	7	7	9	10	10	10
	4	0	0	0	1	5	5
	5	0	0	17	149	277	277
	6	7	12	33	114	152	152
	7	6	6	8	10	11	11

(b) Données 54k, 7 populations.

		Paliers Q					Total
		$Q \geqslant 0.99$	$Q \geqslant 0.85$	$Q \geqslant 0.70$	$Q \geqslant 0.50$	$Q \geqslant 0.30$	
Pop.	1	4	21	30	43	43	43
	2	7	43	53	59	59	59

(c) Données 800k, 2 populations.

Table 7.2 – Nombre d'échantillons assignés à chaque population pour cinq seuils Q.

Données	Nombre de populations	Numéro	Ankole	Ankole Zebu cross	East African Shorthorn Zebu	Nganda	Nkiga	Nsongora	Ntoro	Ntuku	Shorthorn Zebu	Small East African Zebu
54k	K=4	1	4	0	0	4	0	0	0	1	0	1
		2	241	7	0	18	28	10	10	0	11	9
		3	53	3	125	15	0	0	0	0	197	45
		4	0	0	8	0	0	0	0	0	5	9
54k	K=7	1	228	6	0	20	28	10	10	0	15	9
		2	0	0	8	0	0	0	0	0	6	9
		3	10	0	0	0	0	0	0	0	0	0
		4	4	0	0	0	0	0	0	1	0	0
		5	46	2	124	3	0	0	0	0	83	19
		6	0	1	1	14	0	0	0	0	109	27
		7	10	1	0	0	0	0	0	0	0	0
800k	K=2	1	29	0	1	3	4	1	0	0	3	2
		2	9	0	16	1	0	0	0	0	26	7

Table 7.3 – Comparaison entre la structure de population calculée par `Admixture` et les races telles que relevées sur le terrain (indications des éleveurs et expertise du partenaire NextGen local). Les vaches désignées comme ankoles (première colonne du tableau) comprennent par exemple 53 individus assignés au cluster 3_4 qui sont donc génétiquement plus proches des zébus. A l'inverse, tous les *East African Shorthorn Zebu* (3^e colonne) semblent appartenir au cluster des zébus.

7.1.3 Détection de la sélection avec Samβada

Afin de détecter des signatures de sélection naturelle, j'ai appliqué Samβada successivement aux données 54k puis aux données 800k et 800ksub, dans le cadre de modèles univariés pour commencer, puis de modèles multivariés. Les 23 variables environnementales décrites à la table 4.5 ont été incluses dans les modèles univariés et les 15 variables contenues dans la table 4.6 ont été utilisées dans le cadre des modèles multivariés. Les données moléculaires ont été recodées en variables indicatrices avec `RecodePlink`. Chaque SNP étant biallélique, il y a trois valeurs possibles pour le génotype (par ex. AA, AG et GG). Chaque SNP a donc été recodé en trois variables binaires indiquant la présente ou l'absence du génotype en question. Comme chaque individu possède un génotype par SNP, une de ces variables indicatrices vaut 1 et les deux autres valent 0.

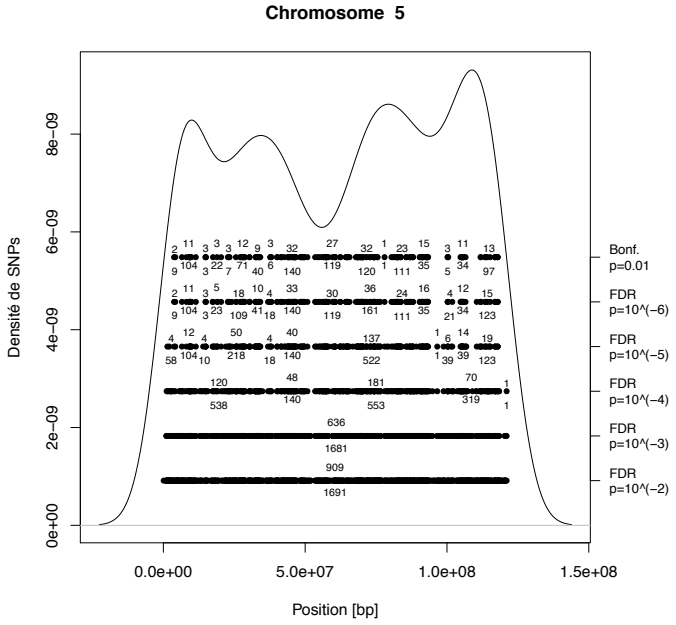
Analyses univariées

Données 54k Les 120'102 marqueurs binaires issus du recodage des 40'034 SNPs ont été utilisés, Samβada a analysé les 2'699'510 modèles en appliquant les tests G et de Wald avec la correction de Bonferroni pour un seuil de significativité de $\alpha = 0.01$. Les 12'782 modèles détectés pour ce seuil correspondent à 2'500 loci potentiellement soumis à la sélection, soit 6,2% des loci analysés. Les 20 premiers modèles selon le score G sont présentés à la table 7.4.

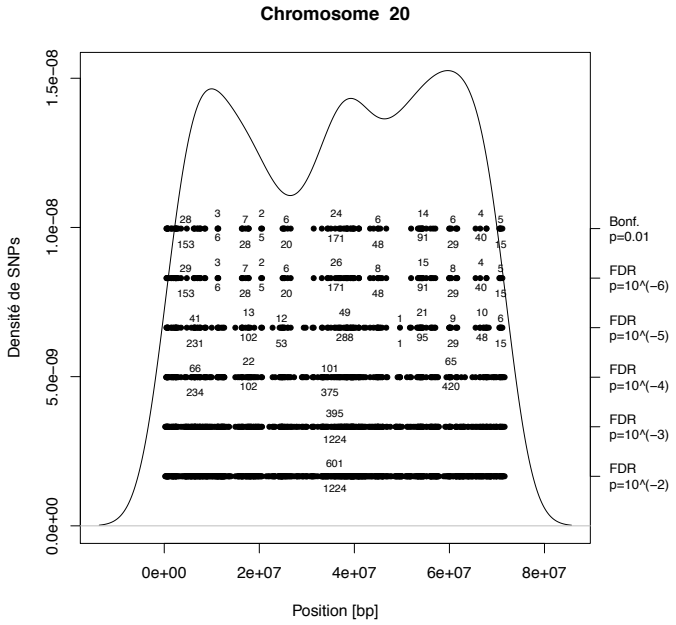
La position des loci détectés sur les chromosomes 5 et 20 est représentée sur la fig. 7.7.

Données 800k Les données recodées en 1'799'094 marqueurs binaires ont servi à calibrer 40,6 millions de modèles. L'application de la correction de Bonferroni, très conservatrice, n'a permis de détecter aucun locus potentiellement sous sélection. Nous avons alors décidé de nous baser sur le score G pour évaluer la significativité des modèles avec la correction de Bonferroni. Dans ce cas, 57 loci, soit 0,01% du total, sont détectés. Les 20 modèles ayant les scores G les plus élevés sont présentés à la table 7.5. La première colonne de la table 7.7 présente le nombre de modèles ayant un score G ou Wald significatif avec la correction de Bonferroni et $\alpha = 0.01$.

D'autre part, deux méthodes de contrôle du taux de faux positifs ont également été employées. La table 7.7 présente les modèles détectés selon Benjamini et Hochberg (1995) et la table 7.9 résume les modèles significatifs selon Storey et Tibshirani (2003) (cf. § 6.1.1.). Ces deux méthodes sélectionnent approximativement le même nombre de modèles pour un même seuil de significativité. Elles ne détectent aucun modèle ayant un score de Wald significatif pour $\alpha = 0,01$. Joost et al. (2007) relèvent à ce propos que le test de Wald peut produire des faux négatifs lorsque l'échantillon est de petite taille. Les SNPs identifiés avec la correction de Bonferroni ou la FDR selon Benjamini et Hochberg (1995) sont dénombrés dans la table 7.8, alors que ceux détectés avec la FDR selon Storey et Tibshirani (2003) sont comptés dans la table 7.10.



(a) Chromosome 5



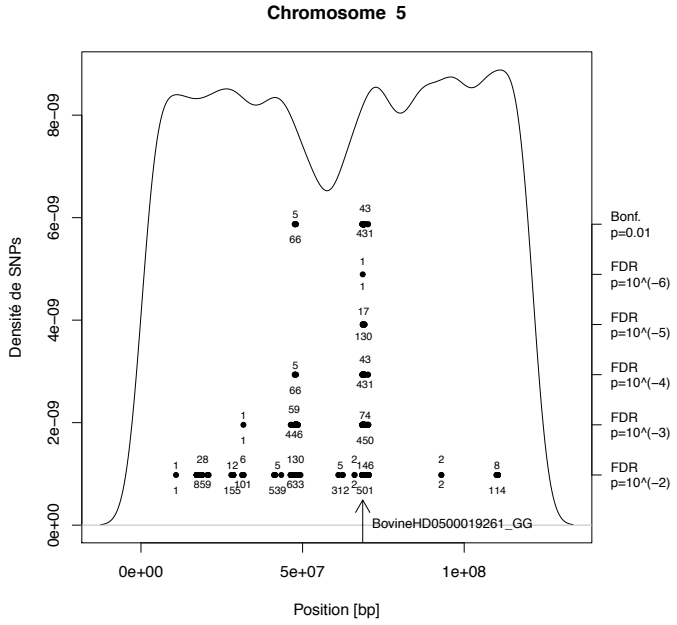
(b) Chromosome 20

Figure 7.7 – Carte des loci détectés sur les chromosomes 5 et 20 avec le jeux de données 54k en fonction du score G. La première marque sur l'axe vertical à droite indique le seuil de significativité $\alpha = 0,01$ (incluant la correction de Bonferroni) et les marques suivantes les seuils de FDR selon Benjamini et Hochberg (1995). L'abscisse donne la position sur le chromosome. L'espacement entre les lignes est arbitraire. Les SNPs sont groupés par cluster, deux loci sont considérés comme voisins si la distance les séparant est inférieure à 2 millions de paires de bases. Chaque cluster est décrit par le nombre de SNPs qu'il contient (dessous) et par le nombre de SNPs détectés (dessus). La courbe représente la densité de SNPs sur la puce ADN, selon l'échelle de l'axe y.

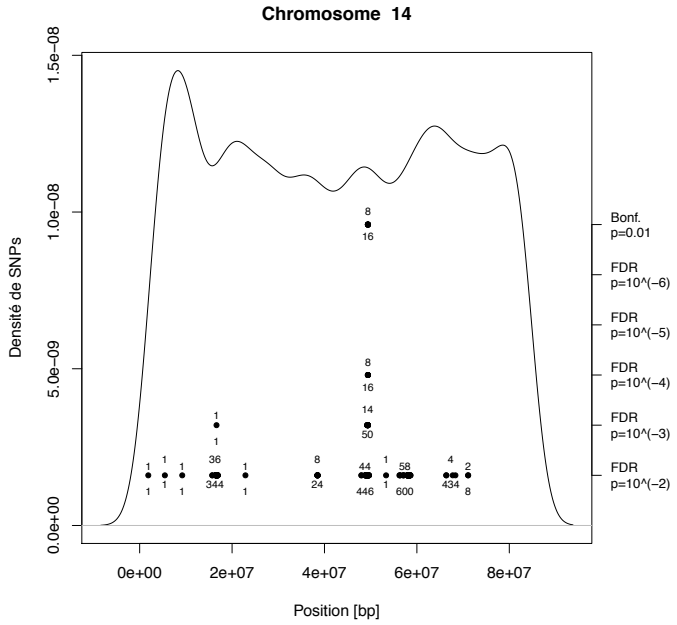
La position des loci détectés sur les chromosomes 5 et 14 est représentée sur la fig. 7.8.

Données 800ksub L'analyse des 91'995 marqueurs binaires issus du recodage des 30'665 SNPs a révélé 12 modèles et 7 loci (0,02 % du total) potentiellement soumis à la sélection (d'après le test *G* avec la correction de Bonferroni, $\alpha = 0.01$). Les modèles et les SNPs sélectionnés avec les deux méthodes de FDR sont dénombrés sur les tables 7.7, 7.8 et 7.9. Les modèles significatifs sont présentés par la table 7.6. La position des loci détectés sur les chromosomes 5 et 6 est représentée sur la fig 7.9.

Samβada détecte de 2'500 loci potentiellement soumis à la sélection dans les données 54k, soit 6,2% du total. Nous savons cependant que les individus ougandais analysés sont répartis entre plusieurs populations, les deux principaux groupes étant les ankoles et les zébus. Cette structure pourrait également expliquer les gradients de fréquences alléliques observés. C'est pourquoi j'ai cherché à savoir si des marqueurs montraient des signatures de sélection à l'intérieur de ces populations.

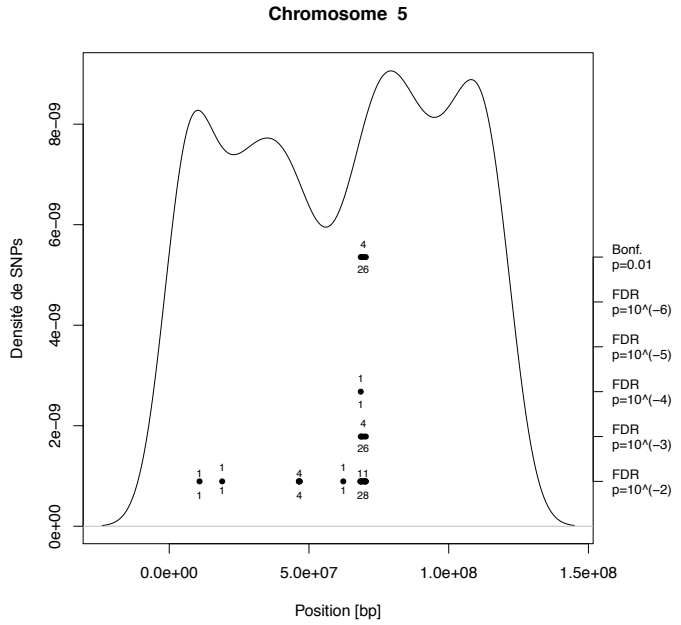


(a) Chromosome 5

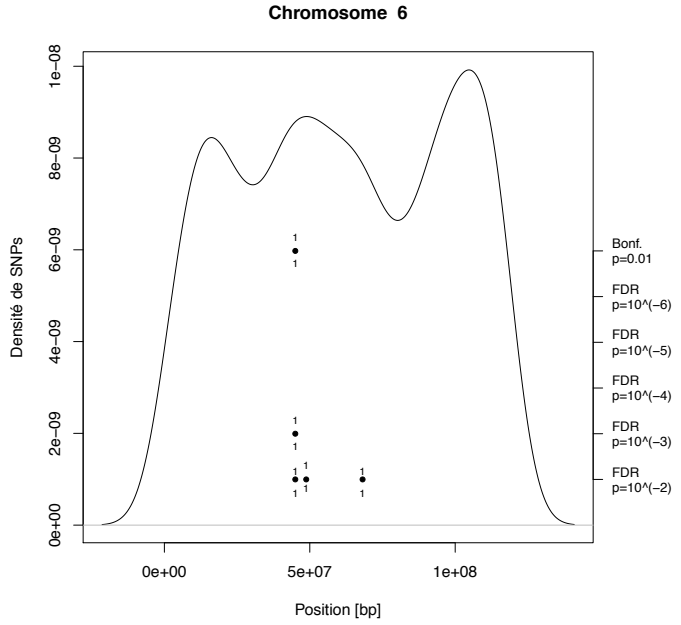


(b) Chromosome 14

Figure 7.8 – Carte des loci détectés sur les chromosomes 5 et 14 avec le jeux de données 800k. La première marque sur l'axe vertical à droite indique le seuil de significativité $\alpha = 0,01$ (incluant la correction de Bonferroni) et les marques suivantes les seuils de FDR selon Benjamini et Hochberg (1995). L'abscisse donne la position sur le chromosome. L'espacement entre les lignes est arbitraire. Les SNPs sont groupés par cluster, deux loci sont considérés comme voisins si la distance les séparant est inférieure à 2 millions de paires de bases. Chaque cluster est décrit par le nombre de SNPs qu'il contient (dessous) et par le nombre de SNPs détectés (dessus). La courbe représente la densité de SNPs sur la puce ADN, selon l'échelle de l'axe y. La flèche indique la position du loci ayant le meilleur score G.



(a) Chromosome 5



(b) Chromosome 6

Figure 7.9 – Carte des loci détectés sur les chromosomes 5 et 6 dans le sous-ensemble de 30k SNPs des données 800k. La première marque sur l'axe vertical à droite indique le seuil de significativité de $\alpha = 0,01$ (incluant la correction de Bonferroni) et les marques suivantes les seuils de FDR selon Benjamini et Hochberg (1995). L'abscisse donne la position sur le chromosome. L'espacement entre les lignes est arbitraire. Les SNPs sont groupés par cluster, deux loci sont considérés comme voisins si la distance les séparant est inférieure à 2 millions de paires de bases. Chaque cluster est décrit par le nombre de SNPs qu'il contient (dessous) et par le nombre de SNPs détectés (dessus). La courbe représente la densité de SNPs sur la puce ADN, selon l'échelle de l'axe y.

Marqueur	Chr.	Position [Mbp]	Env_1	Score G	Score Wald	AIC	BIC	Corrélation
ARS-BFGL-NGS-113888_GG	5	48.30	prec7	208.67	151.70	887.47	910.23	+
ARS-BFGL-NGS-113888_GG	5	48.30	latitude	193.89	146.89	902.25	925.01	+
ARS-BFGL-NGS-113888_GG	5	48.30	prec6	181.19	138.85	914.95	937.71	+
ARS-BFGL-NGS-113888_GG	5	48.30	bio7	179.50	128.90	916.64	939.40	+
Hapmap41074-BTA-73520_AA	5	48.40	prec7	208.53	151.72	890.22	912.98	+
Hapmap41074-BTA-73520_AA	5	48.40	latitude	193.13	146.61	905.62	928.38	+
Hapmap41074-BTA-73520_AA	5	48.40	bio7	180.61	129.73	918.14	940.90	+
Hapmap41074-BTA-73520_AA	5	48.40	prec6	179.99	138.13	918.77	941.53	+
Hapmap41762-BTA-117570_GG	5	18.90	prec7	202.93	148.43	875.92	898.68	+
Hapmap41762-BTA-117570_GG	5	18.90	latitude	186.04	141.99	892.80	915.56	+
Hapmap41762-BTA-117570_GG	5	18.90	prec6	166.51	130.97	912.33	935.09	+
ARS-BFGL-NGS-46098_GG	20	3.00	prec7	200.82	147.60	884.07	906.83	+
ARS-BFGL-NGS-46098_GG	20	3.00	latitude	178.45	138.11	906.44	929.20	+
ARS-BFGL-NGS-46098_GG	20	3.00	prec6	177.87	137.27	907.03	929.78	+
ARS-BFGL-NGS-46098_GG	20	3.00	bio7	167.50	121.71	917.39	940.15	+
ARS-BFGL-NGS-46098_GG	20	3.00	longitude	163.18	130.95	921.72	944.48	+
Hapmap41813-BTA-27442_AA	5	49.00	prec7	179.89	137.52	928.60	951.36	+
Hapmap41813-BTA-27442_AA	5	49.00	latitude	164.71	130.98	943.77	966.53	+
BTA-73516-no-rs_AA	5	48.80	prec7	177.43	136.04	924.35	947.11	+
BTA-73516-no-rs_AA	5	48.80	latitude	161.06	128.61	940.72	963.48	+

Table 7.4 – Modèles univariés ayant les plus hauts scores *G* pour les données 54k. Chaque marqueur est désigné par le nom du locus sur la puce ADN, suivi de l'allèle considéré (par ex. AA, AG ou GG). Ces marqueurs sont issus de trois régions situées sur les chromosomes 5 et 20. Un marqueur peut être associé à plusieurs variables environnementales (cf table 4.5). Une corrélation positive indique que la fréquence du marqueur augmente avec la valeur de la variable environnementale.

Marqueur	Chr.	Position [Mbp]	Env_1	Score G	Score Wald	AIC	BIC	Corrélation
BovineHD0500019261_GG	5	68.60	latitude	64.41	27.17	77.80	92.30	+
BovineHD0500019261_GG	5	68.60	prec7	56.00	26.10	86.21	100.71	+
BovineHD0500019261_GG	5	68.60	bio9	55.05	23.54	87.16	101.66	+
BovineHD0500019261_GG	5	68.60	bio7	54.19	25.69	88.02	102.52	+
BovineHD0500019261_GG	5	68.60	prec2	48.25	24.26	93.96	108.46	-
BovineHD0500019315_GG	5	68.80	bio9	53.17	23.01	90.30	104.80	+
BovineHD0500019259_GG	5	68.60	latitude	52.13	27.17	90.08	104.58	+
BovineHD0500019259_GG	5	68.60	prec7	51.35	25.96	90.86	105.36	+
BovineHD0500019312_AA	5	68.80	bio9	51.30	21.71	93.95	108.45	+
BovineHD0500019393_GG	5	69.20	bio9	49.60	23.99	91.02	105.52	+
BovineHD0500019363_AA	5	69.00	bio9	49.32	20.10	95.10	109.60	+
BovineHD0500019389_GG	5	69.20	bio9	49.22	23.96	91.40	105.90	+
BovineHD0500019295_GG	5	68.70	bio9	49.00	20.97	96.36	110.86	+
BovineHD0500019298_CC	5	68.70	bio9	48.72	20.51	96.33	110.83	+
BovineHD0500019299_GG	5	68.70	bio9	48.72	20.51	96.33	110.83	+
BovineHD0500019362_AA	5	69.00	bio9	48.66	20.29	96.11	110.61	+
BovineHD0500019296_CC	5	68.70	bio9	48.36	19.49	94.52	109.02	+
BovineHD0500019289_GG	5	68.70	bio9	48.36	19.49	94.52	109.02	+
BovineHD0500019297_AA	5	68.70	bio9	48.36	19.49	94.52	109.02	+
BovineHD0500019291_GG	5	68.70	bio9	48.07	19.86	95.92	110.42	+

Table 7.5 – Modèles univariés ayant les plus hauts scores *G* pour les données 800k. Chaque marqueur est désigné par le nom du locus sur la puce ADN suivi de l'allèle considéré (par ex. AA, AG ou GG). Ces marqueurs sont issus de la même région du chromosomes 5. Un marqueur peut être associé à plusieurs variables environnementales (cf table 4.5). Une corrélation positive indique que la fréquence du marqueur augmente avec la valeur de la variable environnementale.

Marqueur	Chr.	Position [Mbp]	Env_1	Score G	Score Wald	AIC	BIC	Corrélation
ARS-BFGL-NGS-111469_AA	5	68.50	prec7	46.74	25.67	98.04	112.54	+
ARS-BFGL-NGS-111469_AA	5	68.50	latitude	41.86	25.88	102.92	117.42	+
ARS-BFGL-NGS-111469_AA	5	68.50	bio7	36.09	22.45	108.68	123.18	+
ARS-BFGL-NGS-111469_AA	5	68.50	bio9	35.53	20.40	109.25	123.75	+
ARS-BFGL-NGS-111469_AA	5	68.50	prec2	34.42	22.48	110.36	124.85	-
BTB-00254199_AA	6	45.00	longitude	40.26	23.54	85.53	100.03	+
Hapmap41127-BTA-101667_AA	20	45.10	longitude	38.92	7.27	25.96	40.46	-
Hapmap57626-rs29027086_AA	5	69.50	latitude	36.51	18.46	62.55	77.05	-
Hapmap27171-BTA-143541_GG	5	68.80	bio9	36.25	17.72	103.45	117.95	+
Hapmap28985-BTA-73836_GG	5	70.30	bio7	35.67	22.26	95.23	109.73	+
Hapmap28985-BTA-73836_GG	5	70.30	bio3	34.57	20.17	96.32	110.82	-
BTB-00316291_GG	7	64.90	prec7	34.73	17.85	57.90	72.40	+

Table 7.6 – Modèles univariés ayant les plus hauts scores *G* pour le sous-ensemble de 30k SNPs des données 800k. Chaque marqueur est désigné par le nom du locus sur la puce ADN, suivi de l'allèle considéré (par ex. AA, AG ou GG). Un marqueur peut être associé à plusieurs variables environnementales (cf table 4.5). Une corrélation positive indique que la fréquence du marqueur augmente avec la valeur de la variable environnementale.

Données	Population	Score	Correction					
			Bonf. $\alpha = 0.01$	FDR $\alpha = 10^{-6}$	FDR $\alpha = 10^{-5}$	FDR $\alpha = 10^{-4}$	FDR $\alpha = 10^{-3}$	FDR $\alpha = 10^{-2}$
54k	tous	G	16426	18370	31836	57748	110762	224234
		Wald	12789	13570	25147	48456	97984	211747
	pop 1	G	1	0	0	0	0	1
		Wald	0	0	0	0	0	0
	pop 2	G	109	13	31	113	452	1913
		Wald	9	0	0	5	9	60
800k	tous	G	104	1	22	105	540	4193
		Wald	0	0	0	0	0	0
	pop 1	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0
	pop 2	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0
800ksub	tous	G	12	0	0	1	12	211
		Wald	0	0	0	0	0	0
	pop 1	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0
	pop 2	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0

Table 7.7 – Nombre de modèles détectés selon leurs scores *G* ou de Wald. La première colonne utilise la correction de Bonferroni et les suivantes la FDR selon Benjamini et Hochberg (1995). L'étiquette « Population » se réfère aux échantillons utilisés : « tous » désigne les analyses comprenant l'ensemble des échantillons, alors que « pop 1 » et « pop 2 » désignent les analyses sur les ankoles et les zébus respectivement.

Données	Population	Score	Correction					
			Bonf. $p = 0.01$	FDR $p = 10^{-6}$	FDR $p = 10^{-5}$	FDR $p = 10^{-4}$	FDR $p = 10^{-3}$	FDR $p = 10^{-2}$
54k	tous	G	2988	3262	4938	7683	12127	19326
		Wald	2500	2598	4190	6789	11034	18531
	pop 1	G	1	0	0	0	0	1
		Wald	0	0	0	0	0	0
	pop 2	G	38	8	13	39	125	538
		Wald	5	0	0	3	5	29
800k	tous	G	57	1	17	57	207	1629
800ksub	tous	G	7	0	0	1	7	76

Table 7.8 – Nombre de SNPs potentiellement soumis à la sélection. Ces loci sont identifiés en fonction des scores *G* et de Wald des modèles les incluant. La première colonne utilise la correction de Bonferroni et les suivantes la FDR selon Benjamini et Hochberg (1995). L'étiquette « Population » se réfère aux échantillons utilisés : « tous » désigne les analyses comprenant l'ensemble des échantillons, alors que « pop 1 » et « pop 2 » désignent les analyses sur les ankoles et les zébus respectivement. Seules les configurations où des SNPs ont été détectés sont reportées ici (cf table 7.7).

Données	Population	Score	Q-valeur					
			0.001	0.005	0.01	0.05	0.1	0.2
54k	tous	G	133536	226597	292485	598688	825845	1193979
		Wald	115068	197826	254209	483902	669234	984685
	pop 1	G	0	0	0	0	339	4116
		Wald	0	0	0	0	0	0
	pop 2	G	556	1360	2328	11440	35185	157323
		Wald	8	29	61	1887	7311	32743
800k	tous	G	565	2551	5152	41642	137128	553136
		Wald	0	0	0	0	0	0
	pop 1	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0
	pop 2	G	0	0	0	0	0	4
		Wald	0	0	0	0	0	0
800ksub	tous	G	20	106	231	1578	5112	22164
		Wald	0	0	0	0	0	0
	pop 1	G	0	0	0	0	0	2
		Wald	0	0	0	0	0	0
	pop 2	G	0	0	0	0	0	0
		Wald	0	0	0	0	0	0

Table 7.9 – Nombre de modèles sélectionnés en contrôlant le taux de faux positifs selon Storey et Tibshirani (2003). La significativité des modèles est testée avec leurs scores *G* ou de Wald. L'étiquette « Population » se réfère aux échantillons utilisés : « tous » désigne les analyses comprenant l'ensemble des échantillons, alors que « pop 1 » et « pop 2 » désignent les analyses sur les ankoles et les zébus respectivement.

Données	Population	Score	Q-valeur					
			0.001	0.005	0.01	0.05	0.1	0.2
54k	tous	G	14002	20477	24761	37111	39120	39909
		Wald	12304	17764	20999	30834	35803	39239
	pop 1	G	0	0	0	0	186	1674
		Wald	0	0	0	0	0	0
	pop 2	G	153	374	660	3646	10650	31196
		Wald	5	16	29	656	2293	8380
800k	tous	G	213	988	2052	15315	45588	144870
	pop 2	G	0	0	0	0	0	2
800ksub	tous	G	10	41	82	601	1747	6196
	pop 1	G	0	0	0	0	0	2

Table 7.10 – Nombre de SNPs potentiellement soumis à la sélection en contrôlant le taux de faux positifs selon Storey et Tibshirani (2003). Ces loci sont identifiés en fonction des scores *G* et de Wald des modèles les incluant. L'étiquette « Population » se réfère aux échantillons utilisés : « tous » désigne les analyses comprenant l'ensemble des échantillons, alors que « pop 1 » et « pop 2 » désignent les analyses sur les ankoles et les zébus respectivement. Concernant les données 800k et 800ksub, seules les configurations où des SNPs ont été détectés sont reportées ici (cf table 7.9).

Analyse par populations

J'ai recherché les SNPs qui présentent une signature de sélection dans les populations identifiées d'ankoles et de zébus. J'ai sélectionné les individus issus de chaque population selon leur plus haut coefficient d'appartenance (cf. p. 96). La population 1 désigne les ankoles et la population 2 les zébus.

Les tables 7.7 et 7.8 dénombrent les modèles et SNPs détectés avec la correction de Bonferroni et la FDR selon Benjamini et Hochberg, alors que les tables 7.9 et 7.10 dénombrent les modèles et loci détectés d'après leur q-valeur calculée selon Storey et Tibshirani. Avec un seuil de significativité $\alpha = 0,01$, seul un marqueur est détecté avec la correction de Bonferroni chez les ankoles (0,002%), alors que 38 marqueurs (0,1%) sont détectés avec le score G et 5 (0,01%) avec le score de Wald chez les zébus (cf table 7.8 première colonne). Ce faible nombre de détections permet de rechercher de plus amples renseignements sur les loci concernés. La vache est une espèce très étudiée et de nombreuses informations sur son génome sont disponibles en ligne. Le site Ensembl [4] fournit une cartographie détaillée des génomes des espèces modèles (Flicek et al., 2013). Les informations reportées dans ce chapitre se réfèrent à la version 73 (septembre 2013). Le site UniProt [5] inventorie les protéines identifiées chez plusieurs espèces (The UniProt Consortium, 2013). La documentation des puces ADN fournit la position précise des SNPs sur le génome, ce qui permet de retrouver la portion de chromosome correspondante dans Ensembl. L'interface graphique du site montre les gènes présents dans cette région avec un lien vers les protéines correspondantes répertoriées dans Uniprot. Ces deux bases de données permettent ainsi de déterminer si un SNP est associé à une fonction métabolique connue.

Le SNP « Hapmap27728-BTA-103839 » détecté chez les ankoles se trouve sur le chromosome 21 et est corrélé à la pente. Ensembl situe ce SNP dans le gène « LRFN5 » qui est annoté *leucine rich repeat and fibronectin type III domain containing 5 precursor*. La fibronectine est une protéine présente dans la matrice extra-cellulaire, elle est impliquée dans la guérison des blessures et le développement embryonaire. Le domaine sus-mentionné est un fragment de la protéine conservé au cours de l'évolution, il est très courant dans le règne animal.

La table 7.11a présente les 9 modèles détectés chez les zébus, avec le score de Wald et un seuil $\alpha = 0,01$. Ces marqueurs sont également détectés au moyen du test G et les trois premiers modèles sont communs aux deux approches. Les trois premiers SNPs se réfèrent à la même région du chromosome 5. Les deux premiers loci se situent à la limite et juste à côté du gène « BT.42818 » décrit comme *DNA-directed RNA polymerase III subunit RPC2*. D'après UniProt [6], la protéine homologue chez l'humain joue un rôle important dans la détection et l'élimination des bactéries intracellulaires et des virus à ADN. Le troisième locus correspond au gène « RFX4-201 » qui est un facteur de transcription [7]. Chez la souris, ce gène est impliqué

4. www.ensembl.org
5. http ://www.uniprot.org
6. http ://www.uniprot.org/uniprot/Q9NW08
7. Protéine nécessaire à l'initiation ou à la régulation de la transcription (copie des régions codantes de l'ADN en molécules ARN qui sont chargées de transmettre l'information dans la cellule).

dans le développement du cerveau. Le quatrième SNP détecté est situé entre deux gènes sur le chromosome 21. Le dernier loci identifié est placé près du gène « NTM » sur le chromosome 29 qui pourrait intervenir dans la croissance et l'adhésion des neurites.

L'analyse séparée des populations ne permet pas de détecter de signatures de sélection dans les jeux de données 800k et 800ksub. En effet, le faible nombre d'individus (43 et 59) limite la puissance du test.

Marqueur	Env_1	Score *G*	Score Wald	Corrélation
Hapmap28985-BTA-73836_GG	bio3	58.35	45.37	-
Hapmap28985-BTA-73836_GG	longitude	48.18	43.06	-
Hapmap28985-BTA-73836_GG	bio7	40.50	35.55	+
BTA-73842-no-rs_GG	bio3	56.92	44.94	-
ARS-BFGL-NGS-106520_AA	bio3	57.73	44.60	-
ARS-BFGL-NGS-106520_AA	longitude	47.39	42.13	-
ARS-BFGL-NGS-106520_AA	bio7	40.34	35.28	+
ARS-BFGL-NGS-22129_GG	longitude	43.66	38.24	-
ARS-BFGL-NGS-402_GG	bio2	42.44	38.11	+

(a) Modèles détectés

chr	nom	dist. gen.	pos
5	BTA-73842-no-rs	81.93	70175036
5	ARS-BFGL-NGS-106520	81.93	70199843
5	Hapmap28985-BTA-73836	81.96	70338965
21	ARS-BFGL-NGS-22129	16.89	19682458
29	ARS-BFGL-NGS-402	0.00	35698376

(b) SNPs détectés

Table 7.11 – Modèles les plus significatifs et SNPs détectés dans la population zébu avec les données 54k. Le test de significativité utilise le score de Wald et la correction de Bonferroni. La distance génétique est proportionnelle à la probabilité qu'une recombinaison (échange d'une portion d'ADN entre chromosomes homologues) ait lieu entre l'extrémité du chromosome et cette position lors de la gamétogenèse.

Comme étape ultérieure, et afin de bénéficier de la capacité de Samβada à traiter des modèles multivariés, nous allons inclure la structure de population — par l'intermédiaire du coefficient d'appartenance — comme variable explicative.

Analyse bivariée

Les modèles bivariés impliquent les 15 variables environnementales décrites à la table 4.6. La variable « ankole » est le coefficient d'appartenance de chaque individu à cette population. Cette nouvelle variable explicative va permettre de représenter la structure de population dans Samβada. En effet, l'analyse de la structure en quatre populations avec `Admixture` pour les données 54k a montré que les clusters 1 et 4 ne comptaient que peu d'individus. Les coefficients d'appartenance correspondants ne fournissent donc que peu d'information sur la structure de population totale. Par conséquent, la connaissance des coefficients d'appartenance aux populations ankole et zébu pour chaque individu suffit donc à décrire l'essentiel de la structure de population. Cette observation est cohérente avec l'analyse des données 800k par `Admixture` où seuls deux clusters sont détectés. En ne considérant que ces populations, les coefficients d'appartenances aux populations ankole et zébu sont complémentaires. La variable « ankole » suffit donc à représenter la structure de population dans Samβada.

Les variables environnementales comprennent donc « ankole » et 14 des 23 variables précédemment utilisées. Ces variables ont été choisies en limitant leurs corrélations par paires pour contrôler l'inflation de la variance dans les modèles bivariés (cf sec. 4.5.2). Les modèles univariés servent de base à l'analyse de la significativité des modèles bivariés et sont donc calculés en premier (cf sec 6.1.1, p. 81). La table 7.12a présente les 10 modèles univariés ayant le plus haut score *G* pour les données 54k. Le premier et le dixième sont deux allèles du même SNP. D'après cet aperçu des modèles univariés, la variable « ankole » est celle qui permet de modéliser le plus précisément la distribution des marqueurs. Six marqueurs de la table 7.12a sont également détectés parmi les premiers modèles univariés de la table 7.4. L'inclusion de la variable « ankole » ne semble pas modifier quels SNPs sont impliqués dans les premiers modèles détectés.

Les dix modèles bivariés ayant les plus haut scores *G* sont résumés sur la table 7.12b. Les six marqueurs détectés figurent également dans la table précédente, la présence de ces marqueurs est ainsi bien expliquée par les modèles uni- et bivariés. En revanche, les variables explicatives des premiers modèles bivariés n'incluent pas « ankole ». Cette différence est discutée à la section 8.2.1 p. 162. Les paires de variables environnementales impliquées dans les modèles bivariés sont souvent la longitude et l'isothermalité, ainsi que les précipitations aux mois d'avril et de mai.

Le SNP « ARS-BFGL-NGS-113888 » est le premier détecté parmi les données 54k (cf tables 7.4 et 7.12). La table 7.14 présente le modèle trivarié détecté pour ce marqueur (allèle GG) [8]. Cet unique modèle a été sélectionné car il possède un score *G* (calculé par rapport à ses parents) ainsi qu'un score de Wald significatifs. Cependant, il convient de comparer les valeurs de l'AIC, qui mesure la vraisemblance d'un modèle en relation avec le nombre de paramètres qu'il inclut. A la lecture des résultats, il faut faire attention au fait que ce modèle trivarié a un AIC

8. Seuls les modèles concernant ce SNP ont été considérés, car le calcul de tous les modèles trivariés aurait pris beaucoup de temps.

juste inférieur à un des modèles bivariés. De plus, l'AIC la plus basse (et donc la meilleure vraisemblance) pour le marqueur « ARS-BFGL-NGS-113888_GG » est fournie par le modèle univarié impliquant la variable « ankole ».

La table 7.13 dénombre les modèles et loci détectés avec les tests *G* et Wald et la correction de Bonferroni (p=0,01). Au total 4'752 loci sont détectés avec les modèles univariés soit 11,9%. La comparaison avec les modèles univariés impliquant 23 variables environnementales montre que deux fois plus de loci sont potentiellement soumis à la sélection naturelle. Cependant, environ la moitié d'entre eux sont détectés uniquement avec la variable « ankole ». Le nombre de modèles détectés par les variables environnementales est donc équivalent au cas précédent. Il y a 37 loci impliqués dans un modèle bivarié significatif (0,09%) et, parmi eux, 3 SNPs (0,007%) sont associés à la variable « ankole ».

La table 7.15 présente les modèles bivariés significatifs ayant un parent significatif qui inclut la variable « ankole ». Ces trois loci sont situés sur le chromosome 5. Ils sont compris parmi les SNPs détectés dans la sous-population zébu (cf. table 7.11b). Leur emplacement sur le génome est décrit à la page 114.

Les données 800k ne présentent pas de modèles bivariés significatifs avec la correction de Bonferroni.

Marqueur	Chr.	Pos. [Mbp]	Env_1	Gscore	WaldScore	AIC	BIC	Corrélation
ARS-BFGL-NGS-113888_GG	5	48.32	ankole	408.48	259.78	687.67	710.42	-
Hapmap41074-BTA-73520_AA	5	48.35	ankole	406.41	258.45	692.34	715.10	-
Hapmap41762-BTA-117570_GG	5	18.94	ankole	355.50	245.32	723.35	746.11	-
BTA-73516-no-rs_AA	5	48.75	ankole	270.63	205.48	831.15	853.91	-
ARS-BFGL-NGS-46098_GG	20	2.95	ankole	266.87	205.39	818.03	840.78	-
Hapmap28985-BTA-73836_CC	5	70.34	ankole	264.35	204.35	812.68	835.43	+
Hapmap41813-BTA-27442_AA	5	49.04	ankole	256.17	196.44	852.32	875.07	-
Hapmap50523-BTA-98407_AA	5	46.74	ankole	244.40	191.84	852.62	875.37	-
BTB-01400776_AA	20	2.70	ankole	228.49	158.50	615.43	638.18	+
ARS-BFGL-NGS-113888_AA	5	48.32	ankole	219.60	110.05	451.50	474.26	+

(a) Modèles univariés

Marqueur	Chr.	Pos. [Mbp]	Env_1	Env_2	Gscore	WaldScore	AIC	BIC	Corr. 1	Corr. 2
Hapmap41074-BTA-73520_AA	5	48.35	prec5	prec4	101.12	88.28	928.18	962.32	+	-
Hapmap41074-BTA-73520_AA	5	48.35	longitude	bio3	76.22	62.14	877.27	911.41	+	-
Hapmap28985-BTA-73836_CC	5	70.34	bio12	prec11	99.97	87.96	932.75	966.89	-	+
Hapmap28985-BTA-73836_CC	5	70.34	bio12	bio15	79.67	68.40	962.24	996.37	-	-
ARS-BFGL-NGS-113888_GG	5	48.32	prec5	prec4	98.45	86.19	925.81	959.95	+	-
ARS-BFGL-NGS-113888_GG	5	48.32	longitude	bio3	76.09	61.88	873.44	907.58	+	-
ARS-BFGL-NGS-113888_GG	5	48.32	bio15	prec5	75.89	67.71	944.63	978.77	+	+
ARS-BFGL-NGS-46098_GG	20	2.95	prec5	prec4	94.67	82.94	906.51	940.65	+	-
Hapmap41813-BTA-27442_AA	5	49.04	prec5	prec4	90.65	80.66	960.47	994.61	+	-
BTA-73516-no-rs_AA	5	48.75	prec5	prec4	81.35	73.15	964.03	998.17	+	-

(b) Modèles bivariés

Table 7.12 – Modèles uni- et bivariés ayant les plus hauts scores G pour les données 54k avec 15 variables environnementales. Chaque marqueur est désigné par le nom du locus sur la puce ADN, suivi de l'allèle considéré. Un marqueur peut être associé à plusieurs variables environnementales (tables 4.5 et 4.6). Une corrélation positive indique que la fréquence du marqueur augmente avec la valeur de la variable environnementale.

		Modèles détectés	Loci détectés
Modèles univariés	Total	12'540	4'752
	Non-détectés avec « ankole »	598	398
	Seulement détectés avec « ankole »	3'098	2'436
	Détectés avec « ankole » et d'autres variables	8'844	1'918
Modèles bivariés	Total	122	37
	Sans « ankole »	118	34
	Avec « ankole »	4	3
	Marqueur détecté et avec « ankole » et avec parent « ankole » significatif	8	3

Table 7.13 – Décompte des modèles et loci détectés dans l'analyse bivariée avec les données 54k (seuil de significativité $\alpha = 0,01$). L'augmentation du nombre de loci identifiés avec des modèles univariés par rapport à l'analyse 54k précédente est due aux nombreux loci qui sont uniquement corrélés à la variable « ankole ». Il y a 2'316 loci détectés avec d'autres variables qu'« ankole », ce qui est comparable au cas précédent. Seuls les modèles bivariés ayant au moins un parent significatif sont recensés dans partie inférieure du tableau. La dernière ligne compte les modèles bivariés incluant la variable « ankole » dont le parent univarié incluant cette variable est aussi significatif.

Env_1	Env_2	Env_3	Gscore	WaldScore	AIC	BIC	Corr 1	Corr 2	Corr 3
bio12	prec10	bio3	45.98	42.58	871.88	917.40	+	-	-

Table 7.14 – Modèle trivarié incluant le marqueur « ARS-BFGL-NGS-113888_GG » parmi les données 54k (seuil de significativité $\alpha = 0,01$). Ce locus est situé sur le chromosome 5 à la position ∼48,32 [Mbp].

Marker	Env_1	Env_2	Gscore	WaldScore	AIC	BIC
Hapmap28985-BTA-73836_GG	bio3	ankole	64.70	48.59	703.84	737.98
ARS-BFGL-NGS-106520_AA	bio3	ankole	53.15	44.25	773.58	807.71
BTA-73842-no-rs_GG	bio3	ankole	47.96	40.98	793.94	827.91
Hapmap28985-BTA-73836_GG	latitude	ankole	40.39	37.86	740.25	774.39

Table 7.15 – Modèles bivariés significatifs dont les parents incluant la variable « ankole » sont aussi significatifs pour les données 54k. Le seuil de significativité est fixé à $\alpha = 0,01$ pour les tests *G* et Wald en utilisant la correction de Bonferroni. Ces loci sont situés sur le chromosome 5.

Temps de traitement

La table 7.16 présente le nombre de marqueurs analysés ainsi que le temps de traitement avec Samβada. Lorsqu'un SNP apparaît uniquement sous la forme de deux allèles, le marqueur binaire correspondant au troisième allèle sera toujours égal à 0. Samβada n'analyse pas ces marqueurs binaires monomorphiques. C'est pourquoi le nombre de marqueurs analysés, qui est utilisé pour calculer la correction de Bonferroni, est inférieur au nombre total de marqueurs. Hormis lors de l'analyse bivariée des données 800k, tous les modèles ont été sauvés sur le disque. Cette sauvegarde intégrale a permis d'appliquer la méthode de FDR selon Storey et Tibshirani qui nécessite de calculer la distribution des p-valeurs. Comme il a fallu enregistrer tous les modèles, les résultats ont dû être filtrés selon leurs scores G et Wald avant d'être chargés dans `R`.

		Nombre marqueurs	Nombre marqueurs monomorphiques	Nombre marqueurs analysés	Temps calculs [h]	Temps tri [m]
Data-54k	Tous animaux	120'102	2'732	117'370	1.15	2.10
	Par populations	120'102	5'650	114'452	0.49	1.87
		120'102	5'211	114'891	0.65	1.93
	Bivariés	120'102	2'732	117'370	8.67	10.83
Data-800k	Tous animaux	1'799'094	35'792	1'763'302	2.86	30.72
	Par populations	1'799'094	87'029	1'712'065	1.67	30.43
		1'799'094	83'335	1'715'759	1.98	30.60
	Bivariés	1'799'094	35'792	1'763'302	18.35	0.90
Data-800k-sub	Tous animaux	92'037	2'473	89'564	0.15	1.47
	Par populations	92'037	5'245	86'792	0.09	1.47
		92'037	5'765	86'272	0.10	1.45
	Bivariés	92'037	2'473	89'564	1.06	8.22

Table 7.16 – Vue d'ensemble du traitement avec Samβada. Le nombre de marqueurs analysés, qui ne compte pas les marqueurs binaires identiques pour tous les individus, est utilisé pour calculer la correction de Bonferroni. Le temps de traitement correspond à la somme des temps de calculs si l'analyse a été distribuée sur plusieurs ordinateurs. Le temps nécessaire pour sélectionner les modèles ayant un score supérieur au seuil choisi est indiqué dans la dernière colonne [9].

Cet aperçu des temps de traitements clôt la présentation des analyses corrélatives réalisées avec Samβada. Les résultats sont discutés à la section 8.2. La seconde approche utilisée pour détecter les signatures de sélection est `BayEnv`.

9. Les temps indiqués ici correspondent à l'analyse de tous les chromosomes, "0", X et Y compris. En effet, ces temps ont été mesurés lors de la première version de l'analyse qui incluait tous les chromosomes. Les décomptes de modèles se rapportent aux résultats présentés dans ce chapitre.

7.1.4 Analyses avec BayEnv

BayEnv est une approche corrélative modélisant les fréquences alléliques au sein de populations tout en tenant compte de l'histoire démographique (Coop et al., 2010). L'analyse se déroule en deux étapes. Un ensemble de loci neutres sert d'abord à estimer les corrélations des fréquences alléliques entre les populations (qui sont supposées être issues de la même population ancestrale). Ensuite la détection des signature de sélection utilise tous les loci ainsi que les variables environnementales. Les corrélations entre les populations permettent de modifier l'hypothèse nulle du test (aucune différence de fréquences entre les populations) pour tenir compte de l'histoire démographique (différences de fréquences dues à la dérive génétique).

Préparation des données

BayEnv modélise les fréquences alléliques par population. Dans notre cas, nous avons quatre populations dans le contexte de l'analyse des données 54k et deux populations dans le cas des jeux de données 800k et 800ksub. D'autre part, BayEnv exige un format de données moléculaires particulier : elles doivent être recodées avec deux lignes par SNP et une population par colonne. Chaque ligne compte le nombre de fois qu'un allèle est présent dans chaque population. Les SNPs étant bialléliques, ils sont codés sur deux lignes. Pour ce faire, j'ai modifié le programme RecodePLINK pour qu'il crée les fichiers de données correspondants. Quand aux variables environnementales, elles doivent être centrées et réduites. Lors qu'il y a deux populations, la standardisation des variables environnementales fait qu'elles ont la même valeur absolue pour les deux populations mais sont de signe opposé. Par conséquent, les facteurs de Bayes calculés par BayEnv seront identiques pour les modèles correspondant à ces variables. Les résultats concernant un marqueur sont donc identiques pour toutes les variables explicatives. Les effets respectifs des variables environnementales ne pourront pas être différenciés.

Principe de fonctionnement

BayEnv commence par estimer les corrélations des fréquences alléliques entre populations sous l'effet de la dérive génétique (cf sec. 6.2.1 p. 87). Il utilise un ensemble de loci neutres pour calibrer son modèle. J'ai donc utilisé les résultats de Samβada pour choisir cet ensemble de SNPs neutres. Le principe est de sélectionner des marqueurs qui ne sont significativement associés à aucune variable environnementale. Ces marqueurs sont vraisemblablement neutres. Dans ce but, j'ai considéré le modèle ayant le plus haut score G pour chaque marqueur (noté $G_{\max}$) : si ce modèle n'est pas significatif, aucun autre modèle ne le sera pour ce marqueur, et il est donc neutre. J'ai ensuite classé les marqueurs en fonction de leur score $G_{\max}$ et j'ai sélectionné les marqueurs ayant les valeurs les plus basses. Nous pouvons observer que le regroupement des modèles par marqueur est nécessaire. Un modèle logistique ayant un score G bas n'implique pas forcément que le marqueur est neutre : il pourrait être significativement

associé à une autre variable environnementale. En fin de compte, les ensembles de loci neutres sélectionnés contiennent 1'000 SNPs pour les jeux 54k et 800ksub et 10'000 SNPs pour le jeu 800k. A titre de comparaison, Coop et al. (2010) utilisent 10k SNPs pour estimer la matrice de covariance pour leur jeu de 640k SNPs (issu du *Human Genome Diversity Project*).

Pour les trois jeux de données, j'ai effectué un million d'itérations [10] pour estimer les matrices de covariance. Le manuel de `BayEnv` indique à ce propos que les matrices de covariances peuvent être « moyennées » afin d'obtenir une meilleure estimation des relations entre populations. Mais comme cette méthode faisait planter le programme avec mes données, j'ai donc choisi en fin de compte une matrice calculée lors d'une des dernières itérations.

Détection des signatures de sélection

La deuxième partie de l'analyse concerne la détection des signatures de sélection. Dans ce cas, `BayEnv` requiert que les SNPs soient recodés comme expliqué plus haut, mais également qu'ils soient placés dans des fichiers séparés. J'ai donc créé un fichier d'entrée par SNP, comprenant deux lignes et deux (ou quatre) colonnes (suivant le nombre de populations). La copie des données est relativement longue : il a fallu 21h30 de traitement pour recoder le jeu de données 800k. J'ai ensuite utilisé les dix premiers marqueurs parmi ces données pour estimer le temps de calcul, soit 31 s. Cela représente un temps de traitement de 22,8 jours sur une machine pour le jeu complet. J'ai donc séparé les SNPs en quatre groupes pour répartir les calculs entre plusieurs machines.

Résultats

`BayEnv` fournit un facteur de Bayes pour chaque paire constituée d'un SNP et d'une variable environnementale. En ce qui concerne la significativité de ces modèles, ces résultats ne sont pas directement convertibles en « p-valeurs ». Dans un premier temps, j'ai utilisé une approche empirique comme suggéré par Coop et al. (2010) pour identifier les loci potentiellement soumis à la sélection. Le principe est d'estimer une distribution empirique et de l'utiliser comme étalon pour sélectionner les loci présentant un facteur de Bayes élevé. Une telle distribution est donnée par les résultats associés aux loci neutres c'est-à-dire ceux qui ont servi à estimer la matrice de covariance. Le traitement est effectué séparément pour chaque variable environnementale : d'abord j'ai extrait les facteurs de Bayes des loci neutres, je les ai triés puis j'ai choisi la 10^{e} valeur la plus élevée sur 1'000 pour les jeux de données 54k et 800ksub, et la 100^{e} valeur la plus élevée pour le jeu 800k. Ces valeurs devraient fournir un seuil approximatif pour lequel 1% des loci neutres ont un score significatif. La distribution exacte des facteurs de Bayes étant inconnue, la correction de Bonferroni et les méthodes de FDR ne peuvent pas être appliquées à ce type de données.

La table 7.17a montre les modèles et les SNPs détectés par `BayEnv` pour un seuil de significati-

10. `BayEnv` utilise une approche de Monte-Carlo par Chaînes de Markov (cf p. 87).

vité de 1% avec cette méthode. Le nombre de SNPs détectés étant très élevé dans les données 54k (64%), j'ai refait la même analyse avec un seuil empirique à 0.1% (table 7.17b). Le nombre de loci détectés dans les données 54k est toujours élevé (49%), ce qui met en doute la validité de cette approche pour identifier les loci soumis à la sélection dans cette étude.

Données	Nombre modèles détectés		Nombre SNPs détectés	
	Observés	Potentiels	Observés	Potentiels
54k	359'252	9'208	25'531	400
800k	93'915	5'997	93'915	5'997
800ksub	723	306	723	306

(a) Seuil empirique à 1%.

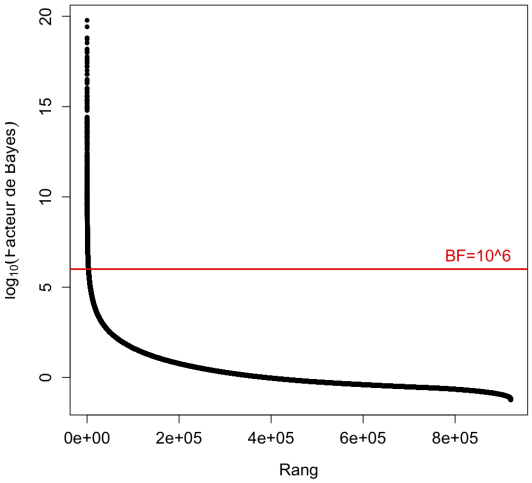
Données	Nombre modèles détectés		Nombre SNPs détectés	
	Observés	Potentiels	Observés	Potentiels
54k	250'564	921	19'561	40
800k	12'749	598	12'749	598
800ksub	116	31	116	31

(b) Seuil empirique à 0,1%.

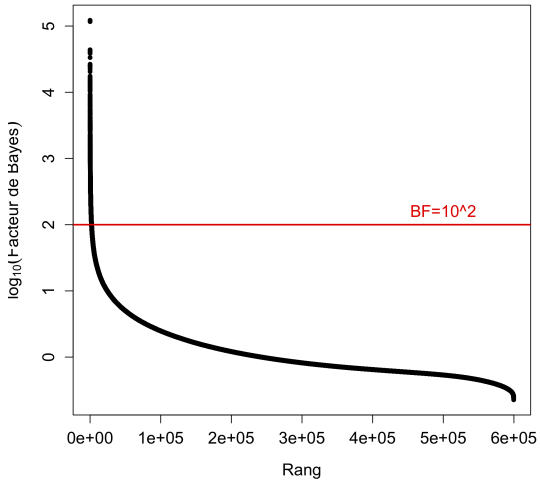
Table 7.17 – Nombre de modèles significatifs et de SNPs détectés avec `BayEnv` selon la première approche. Ces nombres sont identiques pour les données 800k et 800ksub, car l'analyse en deux populations ne permet pas de distinguer les effets des variables environnementales. Le nombre de fausses découvertes potentielles est estimé en supposant que l'hypothèse nulle est vraie pour tous les modèles analysés.

Cette première analyse explique pourquoi j'ai procédé à une deuxième sélection des modèles basée sur les facteurs de Bayes, sans référence aux loci désignés comme neutres. La figure 7.10 présente la distribution des facteurs de Bayes pour les trois jeux de données. J'ai fixé les seuils de significativité en fonction de la valeur maximale des facteurs dans chaque cas, soit à 10^6 pour les données 54k, à 10^2 pour les données 800k et à 10 pour les données 800ksub, Cette sélection des modèles s'appuie sur le fait que le facteur de Bayes, contrairement à une p-valeur, mesure directement la robustesse de l'association entre le marqueur et la variable environnementale. C'est ainsi que la fiabilité des modèles peut être comparée à partir de leur facteur de Bayes. La table 7.18 compte les modèles et les SNPs détectés avec cette approche.

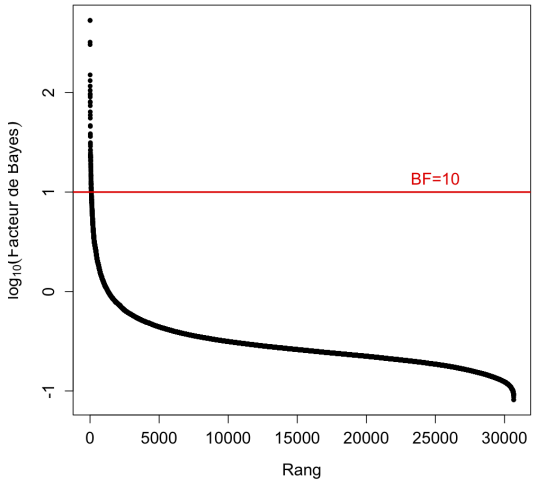
Les résultats de `BayEnv` et Samβada sont comparés sur la table 7.19. Pour les données 54k, les détections de `BayEnv` sont presque toutes comprises dans celles de Samβada (387 sur 400). Si on considère les 100 premiers loci (classés selon leur score G ou leur facteur de Baye), 65 sont communs aux deux méthodes. Les SNPs les plus significatifs sont donc communs à `BayEnv` et Samβada. Pour les données 800k, c'est cette fois-ci `BayEnv` qui englobe une partie des résultats de Samβada (34 SNPs sur 57). Il y a que 2 SNPs communs dans les deux ensembles des 50 SNPs les plus significatifs. Les analyses des données 800ksub ne révèlent pas de loci communs aux deux approches.



(a) Données 54k



(b) Données 800k



(c) Données 800ksub

Figure 7.10 – Distribution des facteurs de Bayes obtenus avec `BayEnv`. L'échelle des facteurs est logarithmique. Les lignes horizontales indiquent les seuils de significativité choisis sur la base des valeurs maximales des facteurs.

Données	Pop.	Seuil	Modèles détectés	SNPs détectés
54k	4	BF $\geqslant 10^6$	3'128	400
800k	2	BF $\geqslant 10^2$	2'083	2'083
800ksub	2	BF $\geqslant 10^1$	91	91

Table 7.18 – Nombre de modèles significatifs et de SNPs détectés avec BayEnv selon la seconde approche. Les seuils de significativité pour les facteurs de Bayes (BF) ont été fixés en fonction de la valeur maximale du facteur pour chaque jeu de données. Ces nombres sont identiques pour les données 800k et 800ksub, car l'analyse en deux populations ne permet pas de distinguer les effets des variables environnementales. Le nombre de fausses découvertes potentielles est estimé en supposant que l'hypothèse nulle est vraie pour tous les modèles analysés.

Données	Nombre SNPs détectés			Comp. premiers SNPs détectés			
	Communs	BayEnv	Samβada	Nb SNPs comparés	Communs	Modèles BayEnv	Modèles Samβada
54k	387	400	2'500	100	65	862	352
800k	34	2'083	57	50	2	50	87
800ksub	0	91	7	-	-	-	-

Table 7.19 – Comparaison des résultats fournis par BayEnv et Samβada. Les modèles BayEnv ont été sélectionnés sur la base de leur facteur de Bayes (voir table 7.18). La partie de gauche compte les SNPs communs ainsi que ceux détectés par chaque programme. La partie de droite compte le nombre de SNPs communs parmi les 100 ou 50 SNPs ayant les scores les plus élevés (facteur de Bayes ou score G). Les deux dernières colonnes indiquent combien de modèles doivent être considérés pour obtenir 100 ou 50 SNPs différents.

Temps de traitement

La table 7.20 présente le temps de traitement avec `BayEnv`. Pour le jeu 800k le temps indiqué dans le tableau est la somme des temps de traitement pour les quatre lots de marqueurs. Le programme utilise parfois plus d'un coeur durant le calcul (100-150% CPU), ce qui accélère le traitement des données. Malgré cela, `BayEnv` est beaucoup plus lent que Samβada qui a analysé les données 54k en un peu plus d'une heure et a traité les données 800k en trois heures. `BayEnv` a calculé pendant 41h dans le premier cas et 26 jours dans le second.

Données	Matrice covariance	Analyse corrélative
54k	40m 37s	41h 17m
800k	21m 11s	629h 13m
		(= 26j 5h)
800ksub	3m 04s	31h 41m

Table 7.20 – Vue d'ensemble du temps de traitement avec `BayEnv`. Les durées indiquent la somme des temps écoulés quand les calculs ont été distribués sur plusieurs ordinateurs [11].

La détection des signatures de sélection naturelle se poursuit avec la deuxième méthode corrélative : `LFMM`.

11. Les temps indiqués ici correspondent à l'analyse de tous les chromosomes, "0", X et Y compris. En effet, ces temps ont été mesurés lors de la première version de l'analyse qui incluait tous les chromosomes. Les décomptes de modèles se rapportent aux résultats présentés dans ce chapitre.

7.1.5 Analyses avec LFMM

Frichot et al. (2013) ont développé les *Latent Factor Mixed Models* (LFMM), une approche corrélative où la structure de population est introduite dans le modèle par l'intermédiaire de variable non-observées. Ce modèle est basé sur les individus et permet de tester la significativité de la corrélation entre la présence des marqueurs et l'environnement, tout en estimant l'influence de ces variables cachées représentant la structure de population.

Préparation des données

LFMM est une méthode basée sur les individus et qui est capable d'analyser des marqueurs bialléliques comme les SNPs. Le fichier de données comprend une ligne par individu et une colonne par SNP, avec comme information le nombre d'occurrences d'un des allèles. Comme le script fourni avec le programme pour la traduction des données depuis le format PLINK n'était pas capable de recoder les données 800k, j'ai adapté l'utilitaire RecodePLINK pour créer les fichiers d'entrée.

Initialement, LFMM ne parvenait pas à charger les données 800k en mémoire et plantait au démarrage. Son auteur, Eric Frichot, m'a indiqué qu'il avait utilisé un cluster pour ses analyses (Frichot et al., 2013). Il m'a alors mis à disposition une version modifiée de l'application que j'ai pu utiliser sur une seule machine.

L'utilisateur doit indiquer à LFMM le nombre de facteurs latents à inclure dans l'analyse. Ces facteurs ne représentent pas directement les populations, ils sont conceptuellement plus proches des axes principaux d'une analyse en composante principale. C'est pourquoi j'ai utilisé un facteur latent ($K = 1$) pour représenter la structure de population des vaches et des zébus.

Résultats

LFMM fournit un score z et une p-valeur pour chaque modèle.

Données	Nombre de modèles significatifs	Nombre de SNPs détectés
54k	303	245
800k	0	0
800ksub	24	14

Table 7.21 – Nombre de modèles significatifs et de SNPs détectés avec LFMM. Le seuil de significativité est fixé à 1% avec la correction de Bonferroni.

Comme le montre la table 7.21, cette approche est conservatrice. La figure 7.11 présente la distribution des p-valeurs pour les trois jeux de données. La configuration b ne permet pas d'appliquer la méthode FDR de Storey et Tibshirani car il faut pouvoir estimer un niveau

moyen de fausses détections pour des p-valeurs proches de 1 comme par exemple sur la figure 6.1 p. 80. Dans les cas a et c, les modèles et loci détectés pour deux seuils de significativité sont indiqués à la table 7.23.

	$q = 0.001$		$q = 0.01$	
Données	Modèles	SNPs	Modèles	SNPs
54k	1'609	1'147	6'494	3'748
800ksub	36	18	114	55

Table 7.22 – Nombre de modèles significatifs et de SNPs détectés avec `LFMM` en appliquant la FDR selon Storey et Tibshirani (2003).

Le contrôle des fausses découvertes selon Benjamini et Hochberg est également applicable aux données 54k et 800ksub car ces résultats ont au moins un modèle significatif avec la correction de Bonferroni. La FDR utilisant un seuil de $\alpha = 1\%$ détecte 1'560 modèles et 1'115 SNPs pour les premières et 31 modèles et 16 SNPs pour les secondes. La table `LFMM` détecte beaucoup plus de SNPs en contrôlant la FDR plutôt qu'en appliquant la correction de Bonferroni, ce qui correspond au comportement de Samβada avec la FDR (cf tab. 7.8 p. 112).

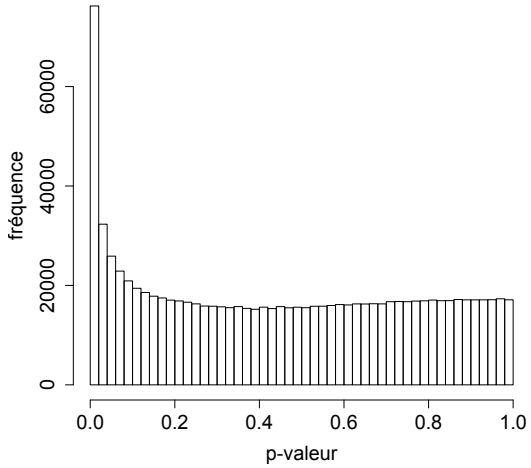
	$q = 0.001$		$q = 0.01$	
Données	Modèles	SNPs	Modèles	SNPs
54k	1'560	1'115	6'208	3'615
800ksub	31	16	114	55

Table 7.23 – Nombre de modèles significatifs et de SNPs détectés avec `LFMM` en appliquant la FDR selon Benjamini et Hochberg (1995).

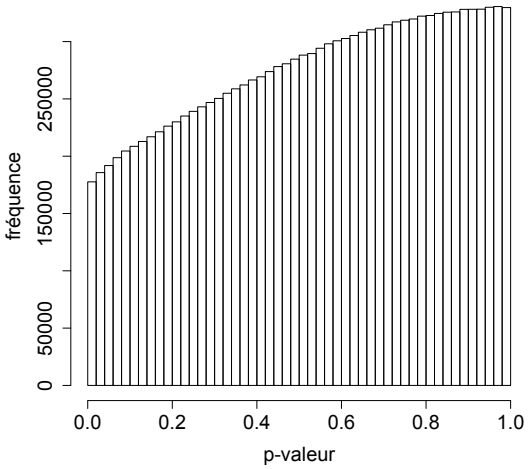
Les résultats de `LFMM` sont comparés avec ceux de Samβada sur la table 7.24. Samβada détecte beaucoup plus de SNPs que `LFMM` dans les données 54k (2'500 contre 245), ce qui explique le nombre de détections communes. J'ai cherché à savoir si Samβada fournit un indice qui permettrait de distinguer les loci qu'il détecte alors que `LFMM` les considère comme neutres. La figure 7.12 présente les distributions de l'AIC et de β'_0 en fonction des détections de Samβada et `LFMM`. β'_0 est le paramètre constant du modèle logistique quand les variables environnementales sont centrées et réduites. `LFMM` semble éliminer les modèles ayant un β'_0 plus grand qu'un seuil légèrement supérieur à 0 ainsi que ceux ayant un petit AIC (à une exception près).

La figure 7.13 compare les résultats des deux approches : elle montre la distribution du score G en fonction du I de Moran global, calculés par Samβada, et met en évidence les SNPs détectés par `LFMM` et par les deux approches simultanément. `LFMM` ne détecte pas les modèles ayant une valeur élevée de ces scores.

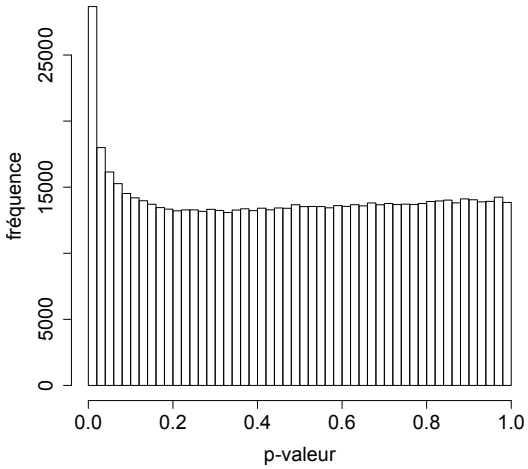
Les deux méthodes identifient un faible nombre SNPs dans les données 800ksub, mais elles en ont cependant quatre en commun (table 7.26). Ces loci communs se sitent sur le chromosome 5, un d'entre eux a également été détecté par Samβada avec les données 54k dans la population



(a) Données 54k

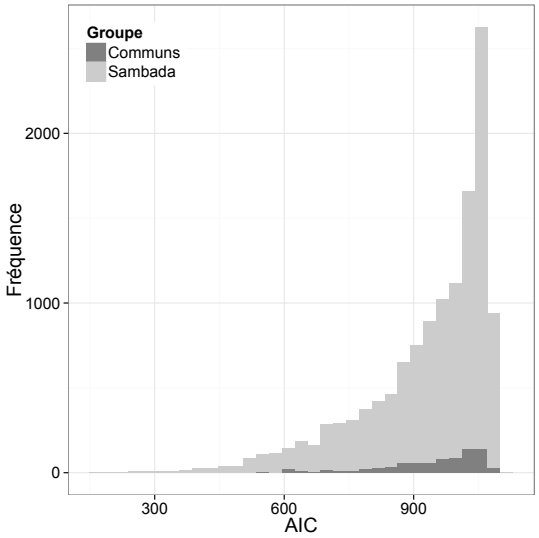


(b) Données 800k

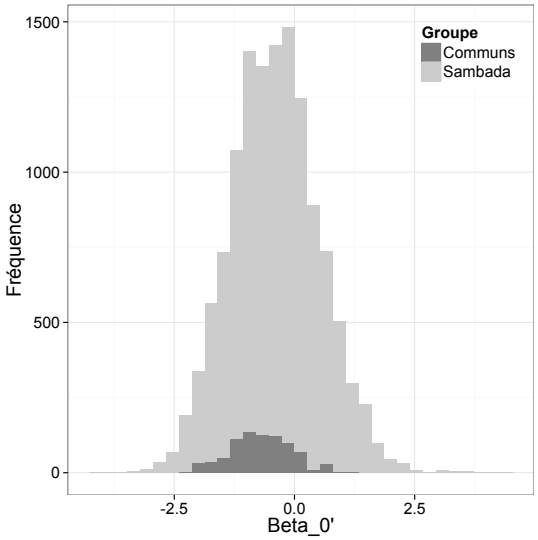


(c) Données 800ksub

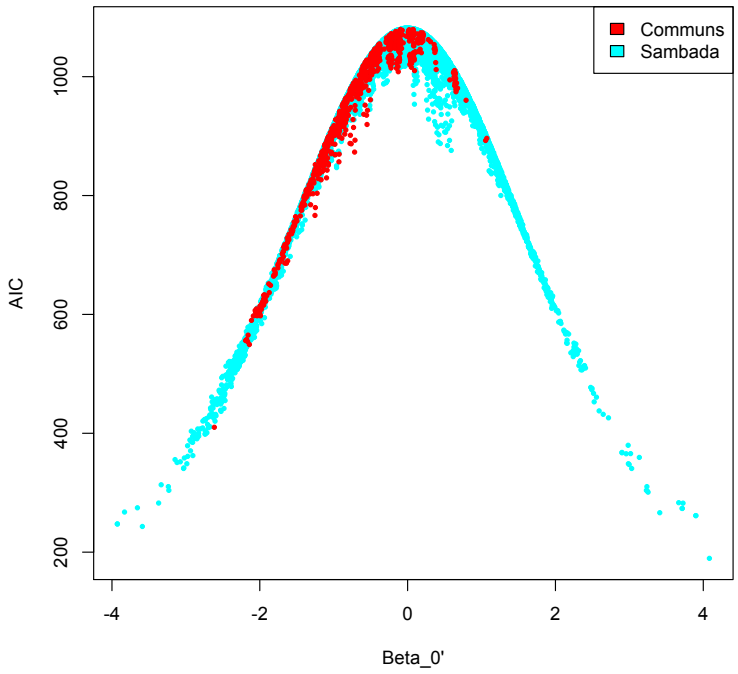
Figure 7.11 – Histogrammes des p-valeurs obtenues avec `LFMM`. Lors d'une sélection de modèles avec le FDR selon Storey et Tibshirani, ces histogrammes sont utilisés pour estimer la proportion de faux positifs.



(a) Distribution de l'AIC



(b) Distribution de β'_0



(c) Distributions combinées de l'AIC et de β'_0

Figure 7.12 – Distributions de l'AIC et de β'_0 en fonction des SNPs détectés par Samβada et `LFMM`. β'_0 est le paramètre constant du modèle si les variables environnementales sont centrées et réduites.

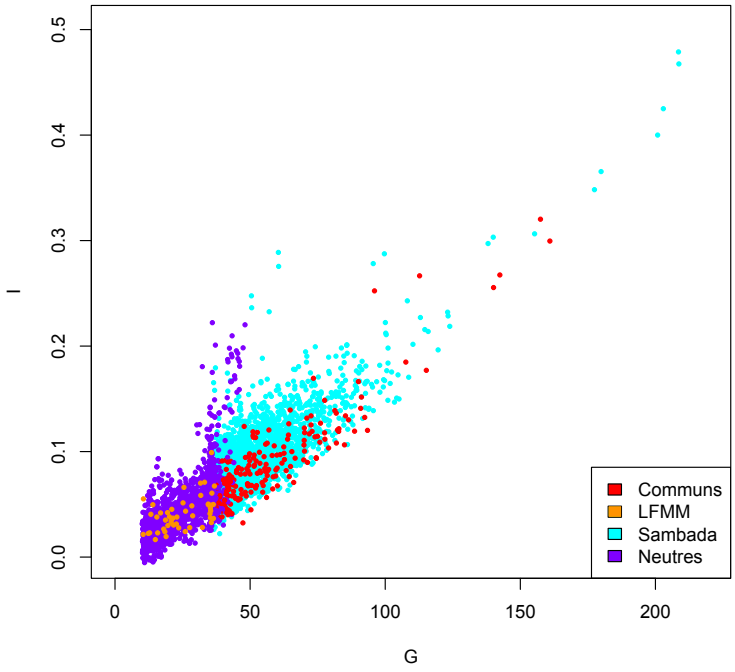


Figure 7.13 – Comparaison des SNPs détectés par Samβada et `LFMM`. Le graphe montre le I de Moran en fonction du score G calculé avec les 20 plus proches voisins. Tous les modèles communs et ceux détectés uniquement par `LFMM` sont représentés. Un échantillon aléatoire des modèles détectés par Samβada et des modèles non-détectés complètent le graphique.

de zébus. Au total, trois loci détectés par LFMM parmi les données 800ksub correspondent à ceux détectés par Samβada avec l'analyse précitée.

	Nombre de modèles détectés			Nombre de loci détectés		
Données	Communs	LFMM	Samβada	Communs	LFMM	Samβada
54k	227	303	12'782	128	245	2'500
800k	0	0	104	0	0	57
800ksub	5	24	12	4	14	7

Table 7.24 – Comparaison des résultats de LFMM et de Samβada pour les trois jeux de données. La partie de gauche compte les modèles communs ainsi que ceux détectés par chaque programme. Les SNPs sont comptés à droite. La p-valeur est fixée 0.01 avant la correction de Bonferroni, pour les deux méthodes.

Temps de traitement

La table 7.25 présente le temps de traitement avec LFMM. Tous les calculs ont été effectués sur une machine, en utilisant plusieurs coeurs. Cette fonctionnalité permet à LFMM de réduire son temps de calcul. LFMM est beaucoup plus rapide que BayEnv mais plus lent que Samβada.

Données	Nombre de coeurs	Temps de traitement	Temps total	Charge du processeur
54k	2	3h 11m	6h 09m	192%
800k	4	16h 02m	58h 05m	355%
800ksub	4	49m	32m	65%

Table 7.25 – Vue d'ensemble du temps de traitement avec LFMM. Tous les calculs ont été effectués sur une machine, en utilisant plusieurs coeurs. La troisième colonne indique le temps écoulé durant le calcul, alors que la quatrième colonne rapporte la somme des temps pour tous les coeurs. La dernière colonne est le quotient des deux précédentes et indique le nombre moyen de coeurs utilisés durant le calcul [12].

La dernière méthode de détection de la sélection est basée sur la génétique des populations.

12. Les temps indiqués ici correspondent à l'analyse de tous les chromosomes, "0", X et Y compris. En effet, ces temps ont été mesurés lors de la première version de l'analyse qui incluait tous les chromosomes. Les décomptes de modèles se rapportent aux résultats présentés dans ce chapitre.

Chr.	Loci	Pos. [Mbp]	Samβada	LFMM	Distance [bp]
4	Hapmap28059-BTA-161846	43.92	0	1	-
5	**ARS-BFGL-NGS-111469**	**68.53**	**1**	**1**	**0**
5	Hapmap50366-BTA-46960	68.61	0	1	85326
5	ARS-BFGL-NGS-33119	68.63	0	1	107042
5	**Hapmap27171-BTA-143541**	**68.82**	**1**	**1**	**0**
5	**Hapmap57626-rs29027086**	**69.55**	**1**	**1**	**0**
5	BTA-73842-no-rs	70.18	0	1	163929
5	ARS-BFGL-NGS-106520	70.20	0	1	139122
5	Hapmap52789-rs29018750	70.26	0	1	80740
5	**Hapmap28985-BTA-73836**	**70.34**	**1**	**1**	**0**
6	BTB-00254199	45.02	1	0	-
7	BTB-00316291	64.89	1	0	-
14	BTA-122374-no-rs	16.44	0	1	-
14	UA-IFASA-5528	16.68	0	1	-
14	Hapmap31734-BTA-129214	49.06	0	1	-
18	ARS-BFGL-NGS-4463	49.17	0	1	-
20	Hapmap41127-BTA-101667	45.14	1	0	-

Table 7.26 – Comparaison des loci détectés par Samβada et LFMM dans les données 800ksub. Le seuil de significativité est $\alpha = 1\%$. Ces approches identifient 7 et 14 SNPs dont quatre concordent. Ces loci communs se situent sur le chromosome 5. La dernière colonne indique la distance (en paires de bases) entre le SNP considéré et le plus proche SNP détecté avec l'autre méthode.

7.1.6 Analyses avec `Arlequin`

`Arlequin` est un logiciel polyvalent d'analyse de données en génétique des populations (Excoffier et Lischer, 2010). Nous décrivons ici uniquement les fonctions liées à la détection de loci sous sélection à partir de la diversité génétique observée entre populations (Excoffier et Lischer, 2011). L'approche utilisée par `Arlequin` est celle de Beaumont et Nichols (1996), elle est basée sur la différenciation entre populations (contrairement aux approches corrélatives utilisées jusqu'ici).

Comme ces analyses ont été réalisés avec d'autres ordinateurs que ceux employés pour les approches corrélatives, les temps de calculs ne sont donc pas comparables à ceux déjà présentés. Nous avons sélectionné des individus clairement assignés à une population en ne prenant que ceux ayant un score d'appartenance égal ou supérieur à 0,85. Cela représente 232 individus du jeu 54k et 64 du jeu 800k. La table 7.27 et la figure 7.14 présentent les échantillons utilisés dans les données 54k. Les zébus y ont proportionnellement moins de représentants que les ankoles. Le filtrage des individus vise à différencier les populations. La distribution de F_{ST} pour des loci neutres a été estimée avec 1'000 itérations pour les données 54k et 100'000 itérations pour les données 800k. Dans le second cas, `Arlequin` ne pouvait pas analyser tous les SNPs simultanément. Nous avons donc distribué les données en lots de 20'000 SNPs.

Population	Individus	Individus analysés	Proportion
1	10	0	0%
2	334	149	44,6%
3	438	73	16,6%
4	22	10	45,5%

Table 7.27 – Décompte des individus considérés par `Arlequin` pour l'analyse des données 54k, ce sont les échantillons avec un coefficient d'appartenance plus grand ou égal à 0,85. La dernière colonne indique quelle proportion de la population est comprise dans l'analyse.

En utilisant la correction de Bonferroni avec $\alpha = 0,01$, `Arlequin` détecte 19 loci dans le jeu de données 54k, dont 6 sont communs à Samβada et 3 concordent avec `LFMM`. Le programme ne détecte aucune signature de sélection significative dans les données 800k.

Les histogrammes des p-valeurs sont présentés sur la fig. 7.15. Ils ne permettent pas d'appliquer la FDR selon Storey et Tibshirani. La FDR de Benjamini et Hochberg détecte également 19 SNPs avec $\alpha = 0,01$ dans les données 54k.

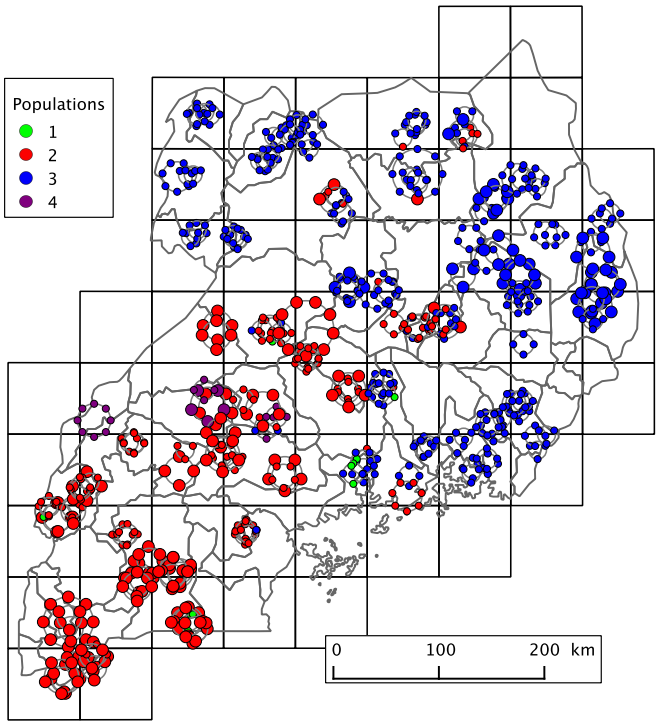
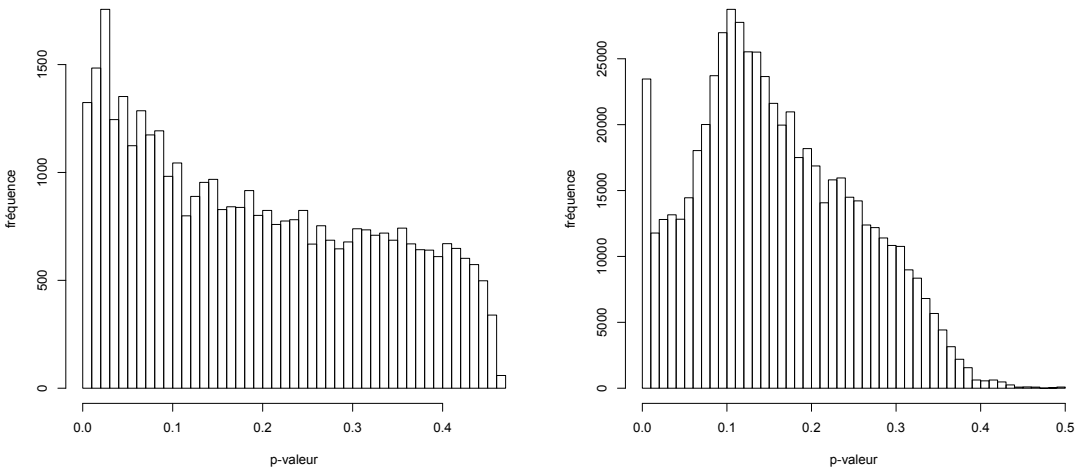


Figure 7.14 – Carte des individus considérés par `Arlequin` pour l'analyse des données 54k. Les grands points indiquent les échantillons ayant un coefficient d'appartenance d'au moins 0,85 qui ont été sélectionnés. Les petits points montrent les individus hybrides qui ont été écartés de l'analyse.



(a) Données 54k

(b) Données 800k

Figure 7.15 – Histogrammes des p-valeurs obtenues avec `Arlequin`.

7.1.7 Comparaison des résultats

Tous les résultats ayant été présentés, nous pouvons procéder à une comparaison globale. La table 7.28 résume le nombre de modèles et de loci détectés par chaque méthode dans les données 54k. Samβada détecte le plus de SNPs potentiellement soumis à la sélection, suivi par BayEnv et LFMM. Arlequin est plus conservateur. La table 7.29 dénombre les SNPs détectés indépendamment par chaque méthode sur chaque chromosome.

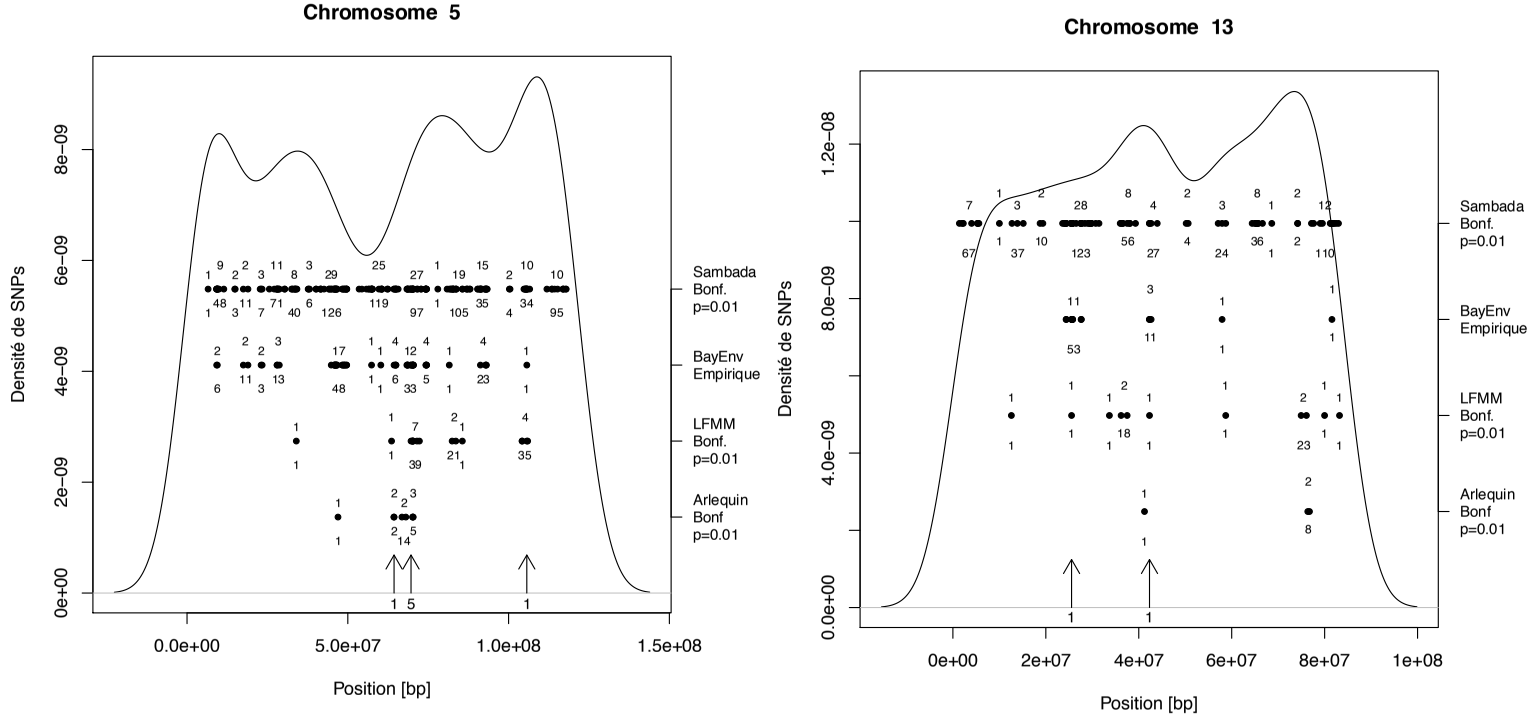
Méthode	Type de détection	Nb modèles	Nb SNPs
Samβada	Bonf. p=0.01	12'782	2'500
BayEnv	Empirique	3'128	400
LFMM	Bonf. p=0.01	303	245
Arlequin	Bonf. p=0.01		19

Table 7.28 – Comparaison du nombre de modèles et de loci détectés par chaque méthode. Samβada, LFMM et Arlequin adaptent le seuil de significativité avec la correction de Bonferroni ($\alpha = 0,01$). BayEnv utilise une approche empirique basée sur le facteur de Bayes.

	1	2	3	4	5	6	7	8	9	10
Samβada Bonf. p=0.01	191	164	134	141	177	101	72	111	97	67
BayEnv Empirique	34	24	22	29	54	22	2	13	14	7
LFMM Bonf. p=0.01	9	14	15	5	16	10	9	8	12	12
Arlequin Bonf p=0.01	1	0	0	0	8	1	0	0	0	0
	11	12	13	14	15	16	17	18	19	20
Samβada Bonf. p=0.01	132	78	81	85	83	50	77	94	68	91
BayEnv Empirique	13	8	16	18	11	4	14	18	10	16
LFMM Bonf. p=0.01	15	5	11	6	10	9	7	6	5	8
Arlequin Bonf p=0.01	1	0	3	0	1	0	0	0	0	2
	21	22	23	24	25	26	27	28	29	
Samβada Bonf. p=0.01	68	68	36	40	22	73	27	28	44	
BayEnv Empirique	14	12	2	1	2	16	2	1	1	
LFMM Bonf. p=0.01	7	7	6	2	1	10	4	4	12	
Arlequin Bonf p=0.01	0	0	0	0	0	0	0	2	0	

Table 7.29 – Nombre de SNPs détectés par Samβada, BayEnv, LFMM et Arlequin sur chaque chromosome. Samβada, LFMM et Arlequin adaptent le seuil de significativité avec la correction de Bonferroni ($\alpha = 0,01$). BayEnv utilise une approche empirique basée sur le facteur de Bayes.

La table 7.30 compare les SNPs détectés par une combinaison de plusieurs méthodes sur chaque chromosome. La plupart des SNPs sont identifiés par une ou deux méthodes, généralement Samβada et BayEnv. Les chromosomes 2, 3, 4, 9, 11, 13, 15, 18, 21 et 22 possèdent des loci détectés par trois méthodes, alors que le chromosome 5 contient des SNPs sélectionnés par trois et même quatre méthodes. La figure 7.16 illustre les régions détectées sur les chromosomes 5 et 13.



(a) Chromosome 5. Les flèches indiquent les régions détectées par trois ou quatre méthodes avec le nombre de loci concernés. Les loci communs à toutes les méthodes se trouvent dans le groupe du centre.

(b) Chromosome 13. Les flèches indiquent les régions détectées par trois méthodes avec le nombre de loci concernés.

Figure 7.16 – Carte des loci détectés sur les chromosomes 5 et 13 par Samβada, `BayEnv`, `LFMM` et `Arlequin` pour les données 54k. L'abscisse donne la position sur le chromosome. L'espacement entre les lignes est arbitraire. Les SNPs sont groupés par cluster, deux loci sont considérés comme voisins si la distance les séparant est inférieure à 2 millions de paires de bases. Chaque cluster est décrit par le nombre de SNPs qu'il contient (dessous) et par le nombre de SNPs détectés (dessus). La courbe représente la densité de SNPs sur la puce ADN, selon l'échelle de l'axe *y*.

Chromosome		1	2	3	4	5	6	7	8	9	10
Nombre de détections	1 méthode	159	135	114	112	125	84	71	98	85	60
	2 méthodes	38	32	27	30	53	25	6	17	16	13
	3 méthodes	0	1	1	1	4	0	0	0	2	0
	4 méthodes	0	0	0	0	3	0	0	0	0	0
Chromosome		11	12	13	14	15	16	17	18	19	20
Nombre de détections	1 méthode	123	73	73	63	74	47	64	77	59	77
	2 méthodes	16	9	16	23	14	8	17	16	12	20
	3 méthodes	2	0	2	0	1	0	0	3	0	0
	4 méthodes	0	0	0	0	0	0	0	0	0	0
Chromosome		21	22	23	24	25	26	27	28	29	
Nombre de détections	1 méthode	56	61	36	39	21	59	27	31	45	
	2 méthodes	15	10	4	2	2	20	3	2	6	
	3 méthodes	1	2	0	0	0	0	0	0	0	
	4 méthodes	0	0	0	0	0	0	0	0	0	

Table 7.30 – Décompte des SNPs communs détectés par Samβada, `BayEnv`, `LFMM` et `Arlequin` sur chaque chromosome avec les données 54k. Seules les correspondances exactes sont reportées. Samβada, `LFMM` et `Arlequin` adaptent le seuil de significativité avec la correction de Bonferroni ($\alpha = 0{,}01$). `BayEnv` utilise une approche basée sur le facteur de Bayes.

Les trois SNPs identifiés par toutes les approches sur le chromosome 5 sont réunis à la table 7.31. Ce sont les mêmes que ceux détectés par Samβada chez les zébus (table 7.11b) ainsi qu'avec les modèles bivariés incluant la structure de population (table 7.15). Les gènes associés ont été présentés à la sec. 7.1.3 p. 114.

Chr.	Loci	Dist. génétique [cMorgan]	Pos. [Mbp]
5	BTA-73842-no-rs	81.93	70.18
5	ARS-BFGL-NGS-106520	81.93	70.20
5	Hapmap28985-BTA-73836	81.96	70.34

Table 7.31 – Loci détectés par Samβada, `BayEnv`, `LFMM` et `Arlequin` dans les données 54k.

Dans le but de comprendre les comportements respectifs des méthodes, comparons les détections en fonction des résultats de Samβada. La table 7.32 présente les SNPs correspondant aux modèles détectés par Samβada avec les plus hauts scores *G*. La partie droite du tableau indique quelles méthodes ont identifié ces SNPs et le nombre de détections total. Les SNPs auxquels Samβada attribue les meilleurs scores *G* sont également détectés par `BayEnv` alors qu'`Arlequin` et `LFMM` ne les détectent pas. Les SNPs identifiés par les quatres méthodes occupent les positions 7, 9 et 10 dans le tableau. Le premier SNP identifié par Samβada, `BayEnv` et `LFMM` mais par `Arlequin` se trouve en 18^{e} place.

Toujours dans le but de mieux comprendre les relations entre les approches, la section suivante compare les distributions spatiales de trois de ces SNPs.

	Loci	Chr.	Pos. [Mbp]	Samβada	BayEnv	LFMM	Arlequin	Détections
1	ARS-BFGL-NGS-113888	5	48.32	1	1	0	0	2
2	Hapmap41074-BTA-73520	5	48.35	1	1	0	0	2
3	Hapmap41762-BTA-117570	5	18.94	1	1	0	0	2
4	ARS-BFGL-NGS-46098	20	2.95	1	1	0	0	2
5	Hapmap41813-BTA-27442	5	49.04	1	1	0	0	2
6	BTA-73516-no-rs	5	48.75	1	1	0	0	2
7	**Hapmap28985-BTA-73836**	**5**	**70.34**	**1**	**1**	**1**	**1**	**4**
8	Hapmap31863-BTA-27454	5	48.99	1	1	0	0	2
9	**ARS-BFGL-NGS-106520**	**5**	**70.20**	**1**	**1**	**1**	**1**	**4**
10	**BTA-73842-no-rs**	**5**	**70.18**	**1**	**1**	**1**	**1**	**4**
11	Hapmap50523-BTA-98407	5	46.74	1	1	0	0	2
12	BTB-01400776	20	2.70	1	1	0	0	2
13	Hapmap23956-BTA-36867	15	47.20	1	1	0	0	2
14	ARS-BFGL-NGS-10586	2	128.64	1	1	0	0	2
15	ARS-BFGL-NGS-43694	5	49.65	1	1	0	0	2
16	BTA-122374-no-rs	14	16.44	1	1	0	0	2
17	BTB-01356178	20	2.49	1	1	0	0	2
18	ARS-BFGL-NGS-94862	11	103.53	1	1	1	0	3
19	BTA-108359-no-rs	14	16.31	1	1	0	0	2
20	ARS-BFGL-NGS-15960	5	28.02	1	1	0	0	2

Table 7.32 – Liste des SNPs détectés par Samβada correspondant aux modèles ayant les plus hauts scores *G* pour les données 54k. Les loci sont identifiés par leur nom, leur chromosome et la position qu'ils y occupent, en millions de paires de bases. Les colonnes suivantes indiquent quelles méthodes les ont détectés et la dernière indique le nombre de ces détections. Les loci en caractères gras sont les découvertes communes aux quatre méthodes.

7.1.8 Autocorrélation spatiale

Les statistiques spatiales nous permettent de tirer profit de la position des individus afin de caractériser la distribution des marqueurs dans l'espace. La mesure de l'autocorrélation spatiale nous permet de déterminer si la présence d'un marqueur chez un individu dépend de sa présence chez ses voisins. Elle nous permet aussi de déterminer où cette relation est significative.

J'ai choisi de comparer des marqueurs issus du jeu de données 54k car toutes les méthodes y détectent des signatures de sélection naturelle. Cette section compare le comportement spatial de trois SNPs issus de la table 7.32. Cette table présente les SNPs que Samβada détecte avec les plus hauts scores *G* dans les données 54k. Contrairement à ce qu'on aurait pu penser, les premiers SNPs sont également détectés par `BayEnv`. Je compare ici le premier SNP de la table, le 7^{e} (qui est la première détection commune aux quatre approches) et le 18^{e} (qu'`Arlequin` ne détecte pas).

Les allèles considérés par Samβada sont notés entre parenthèses.

ARS-BFGL-NGS-113888 (GG) détecté par Samβada et `BayEnv` ;

Hapmap28985-BTA-73836 (GG) premier SNP détecté par toutes les méthodes ;

ARS-BFGL-NGS-94862 (AA) détecté par Samβada `BayEnv` et `LFMM` mais pas par `Arlequin`.

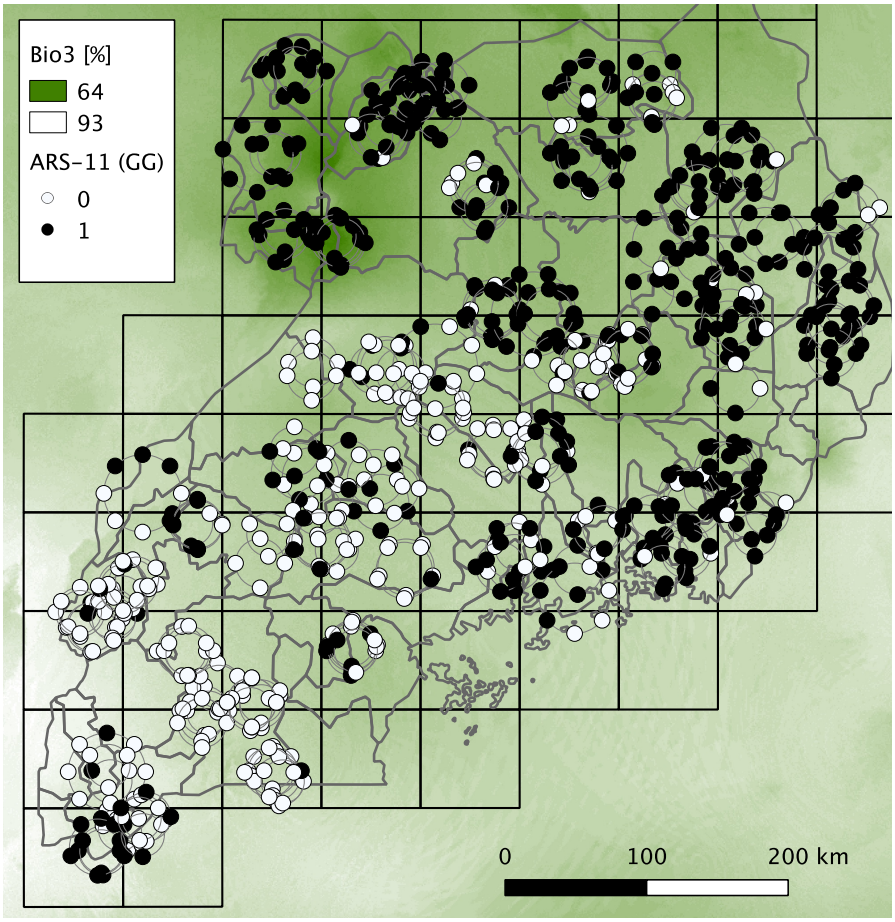
Je renomme ces SNPs dans le corps du texte pour simplifier la notation :

ARS-11 ARS-BFGL-NGS-113888 (GG) ;

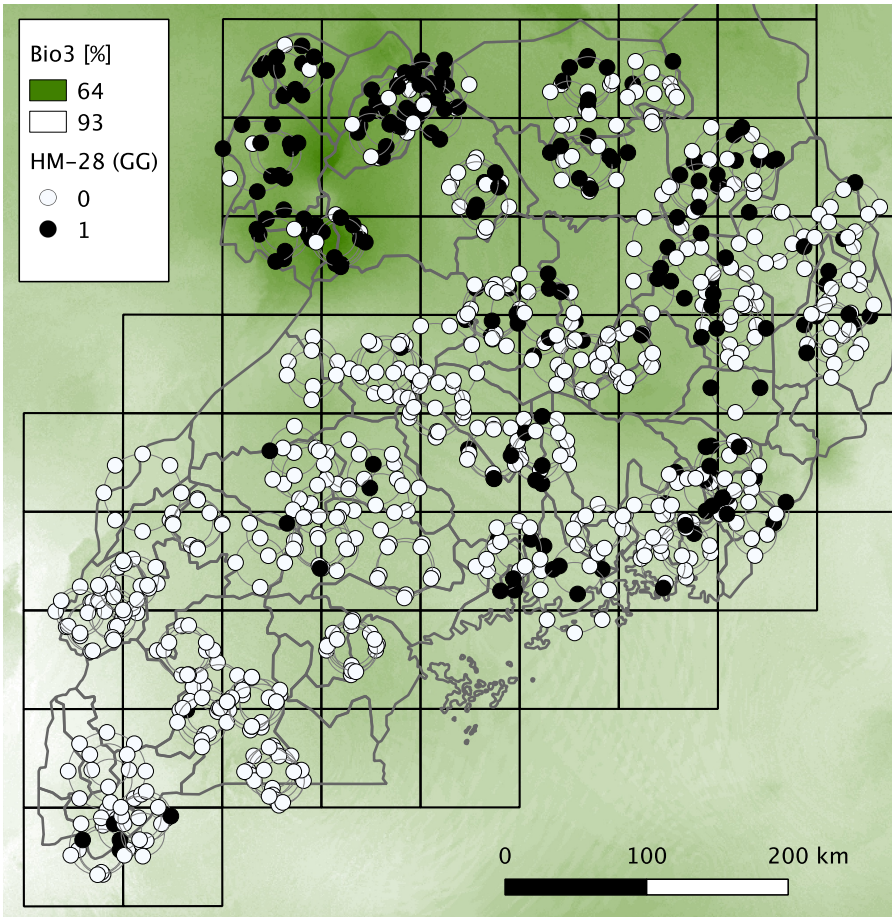
HM-28 Hapmap28985-BTA-73836 (GG) ;

ARS-94 ARS-BFGL-NGS-94862 (AA).

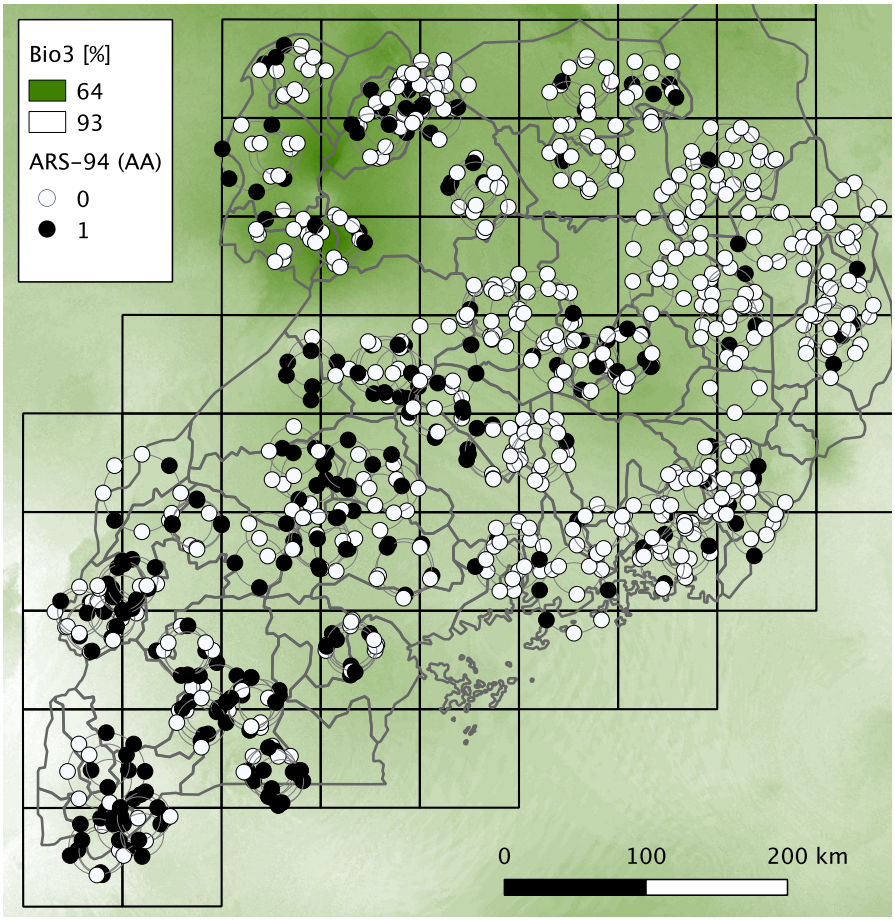
La distribution spatiale de ces marqueurs est illustrée sur la fig. 7.17. Comme plusieurs individus ont été échantillonnés dans chaque ferme, les cartes de cette section les représentent sur des cercles entourant la position géographique de la ferme. Le marqueur ARS-11 est très présent dans le nord et l'est du pays avec quelques occurrences dans le centre et le sud-ouest ; le marqueur HM-28 est principalement présent dans le nord-est, il apparaît également dans l'est et parfois au sud-ouest ; le marqueur ARS-94 est quant à lui principalement présent dans le sud-ouest, il apparaît régulièrement au centre et au nord-ouest et parfois à l'est du pays. Le fond de carte représente l'isothermalité fournie par *WorldClim* (Bio3, table 4.2 p. 32). Cette variable est souvent présente dans les modèles bivariés significatifs tenant compte de la structure de population (table 7.15). L'isothermalité est le rapport entre l'étendue des températures mensuelles moyennes et l'étendue des températures annuelles. Une petite valeur de l'isothermalité indique que l'intervalle de température varie beaucoup d'un mois à l'autre.



(a) ARS-11



(b) HM-28

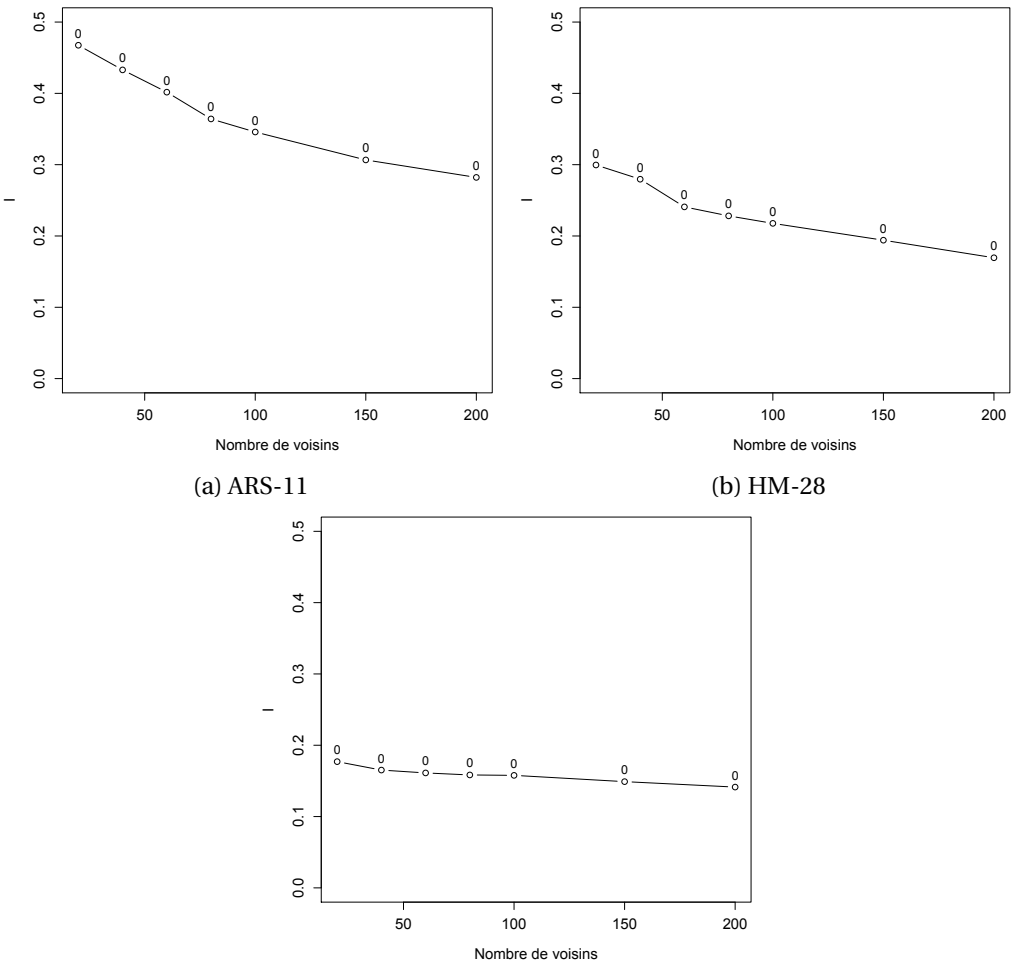


(c) ARS-94

Figure 7.17 – Distribution spatiale des marqueurs ARS-11, HM-28 et ARS-94 issus des données 54k. Les points noirs indiquent la présence du marqueur et les points blancs son absence. Le fond de carte représente l'isothermalité (cf table 4.2 et corps du texte). Les valeurs basses de l'isothermalité sont représentées en vert foncé, car elles correspondent aux régions où la température varie le plus d'un mois à l'autre.

Autocorrélation spatiale globale

Dans cette section, l'autocorrélation spatiale est mesurée par le I de Moran. Toutes les cartes utilisent une pondération basée sur les 20 plus proches voisins. L'autocorrélation spatiale globale de ces marqueurs est résumée par les corrélogrammes de la fig. 7.18. Ces figures montrent que l'autocorrélation décroît quand le nombre de voisins augmente. En effet, plus le voisinage d'un point est grand, plus la moyenne des valeurs dans ce voisinage se rapproche de la moyenne globale. C'est pourquoi on s'attend à voir l'autocorrélation baisser quand le voisinage grandit. D'après la fig. 7.18, l'indice global de Moran ne permet pas de discriminer précisément les loci identifiés par Samβada mais rejetés par les autres méthodes. Les loci HM-28 et ARS-94 présentent une autocorrélation globale assez faible, ce qui correspond à l'observation que LFMM rejette les marqueurs les plus autocorrélés (voir fig. 7.13 p. 131).



(a) ARS-11

(b) HM-28

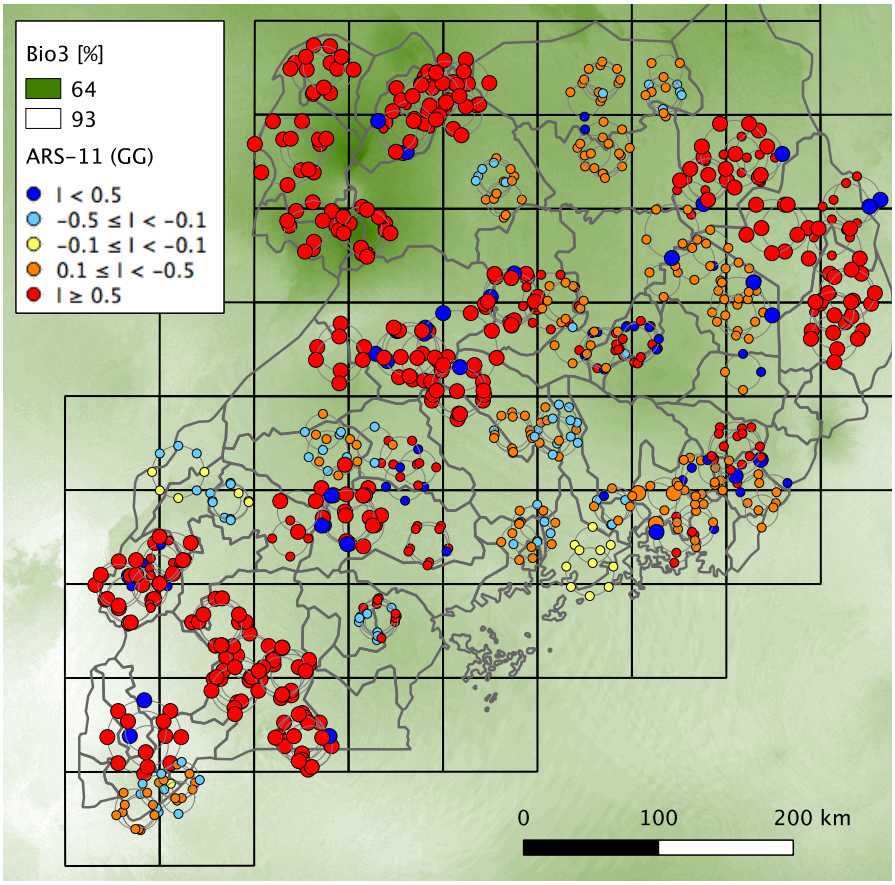
(c) ARS-94

Figure 7.18 – Corrélogrammes des marqueurs ARS-11, HM-28 et ARS-94 issus des données 54k. Ces corrélogrammes ont été calculés avec le I de Moran global et des pondérations basées sur le nombre de plus proches voisins. Les étiquettes indiquent si l'autocorrélation spatiale est significative : une valeur de 0 représente une p-valeur de 0.001 calculée par des permutations.

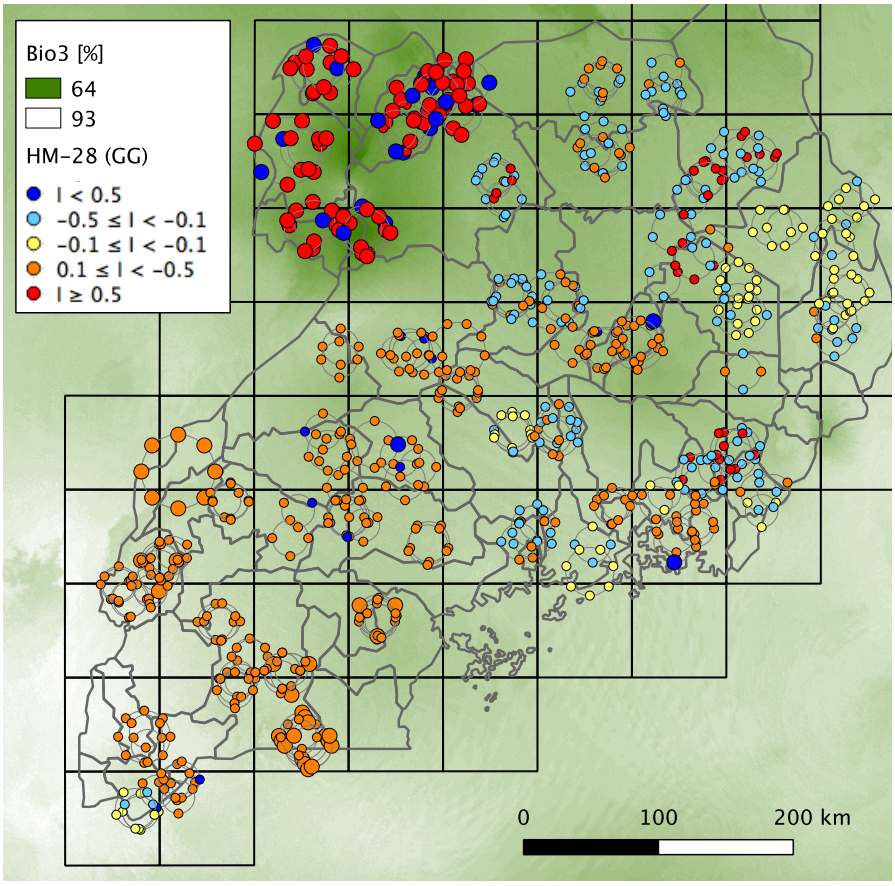
Autocorrélation spatiale locale

La figure 7.19 présente la valeur l'autocorrélation spatiale locale pour les trois marqueurs. Ces marqueurs présentent en général une autocorrélation positive (points rouges) avec quelques points à l'autocorrélation négative (en bleu) : par exemple, ARS-94 est couramment présent au sud-ouest, les points où il est absent dans cette région sont en bleu, car ils ne ressemblent pas à leurs voisins. L'autocorrélation locale du marqueur ARS-11 est significative au nord-ouest, dans l'est, au centre et au sud-ouest du pays, soit sur une portion importante du territoire. Lorsque l'autocorrélation est moins prononcée (en orange et bleu clair), elle n'est généralement pas significative (petits points). Peu de points sont indépendants de leur voisinage et, s'ils existent, ils se situent surtout autour et un peu au nord de Kampala [13] (points jaunes). L'autocorrélation locale de HM-28 et ARS-94 est très différente de celle du premier marqueur (fig. 7.19, b et c). L'intensité est plus faible (points oranges et bleu clairs majoritaires) et elle n'est pas significative sur la presque-totalité du territoire. Les seules régions où l'autocorrélation est forte et significative sont le nord-ouest pour HM-28 et le sud-ouest pour ARS-94, ce qui correspond aux zones où ces marqueurs sont les plus présents. Concernant HM-28, le nord-ouest est également la zone où l'isothermalité est la plus basse, c'est-à-dire où les températures varient le plus d'un mois à l'autre. HM-28 a aussi plus de points indépendants de leur voisinage (en jaune), principalement dans l'est du pays. Ce résultat corrobore l'observation du I de Moran global (p. 143) : les marqueurs HM-28 et ARS-94 sont moins autocorrélés que ARS-11 (indices globaux $I = 0,30$ et $I = 0,18$ contre $I = 0,47$).

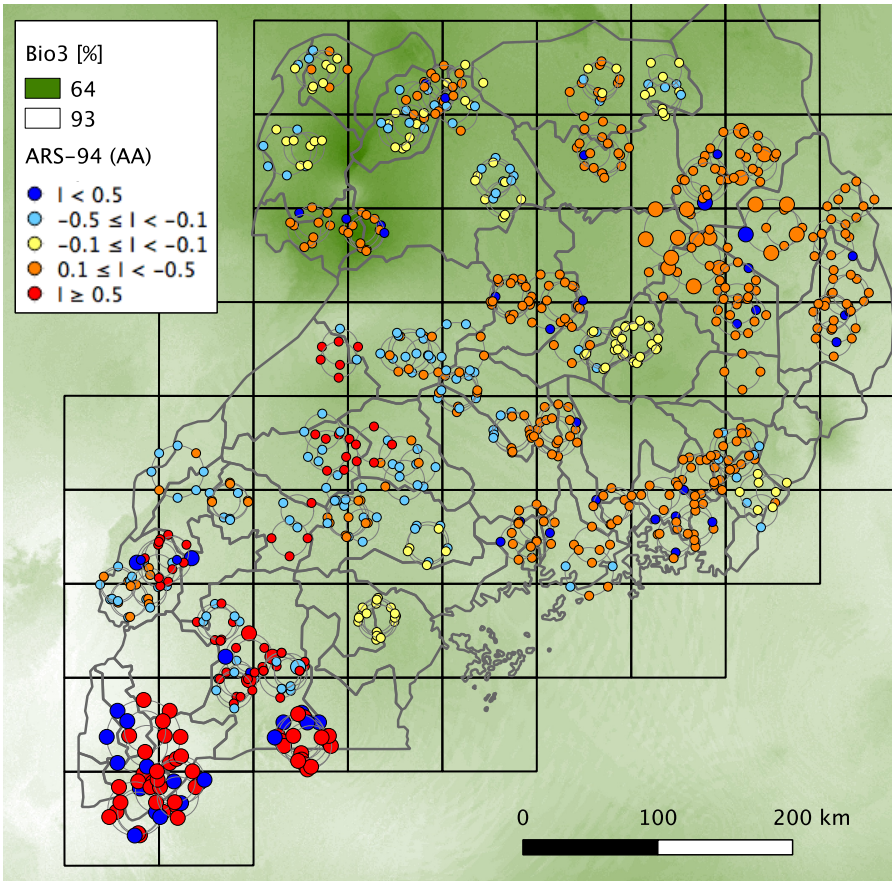
13. La capitale de l'Ouganda est située au sud-est du pays au bord du lac Victoria (voir carte générale p. 50).



(a) ARS-11



(b) HM-28



(c) ARS-94

Figure 7.19 – Indices d'autocorrélation spatiale locale des marqueurs ARS-11, HM-28 et ARS-94 issus des données 54k. La pondération est basée sur les 20 plus proches voisins. Les marqueurs rouges ont tendance à ressembler à leur voisinage alors que les bleus en diffèrent. Les marqueurs jaunes sont indépendants de leurs voisins. Les petits point signalent les valeurs non significatives ($p > 0.001$). Le fond de carte représente l'isothermalité (cf table 4.2 et corps du texte).

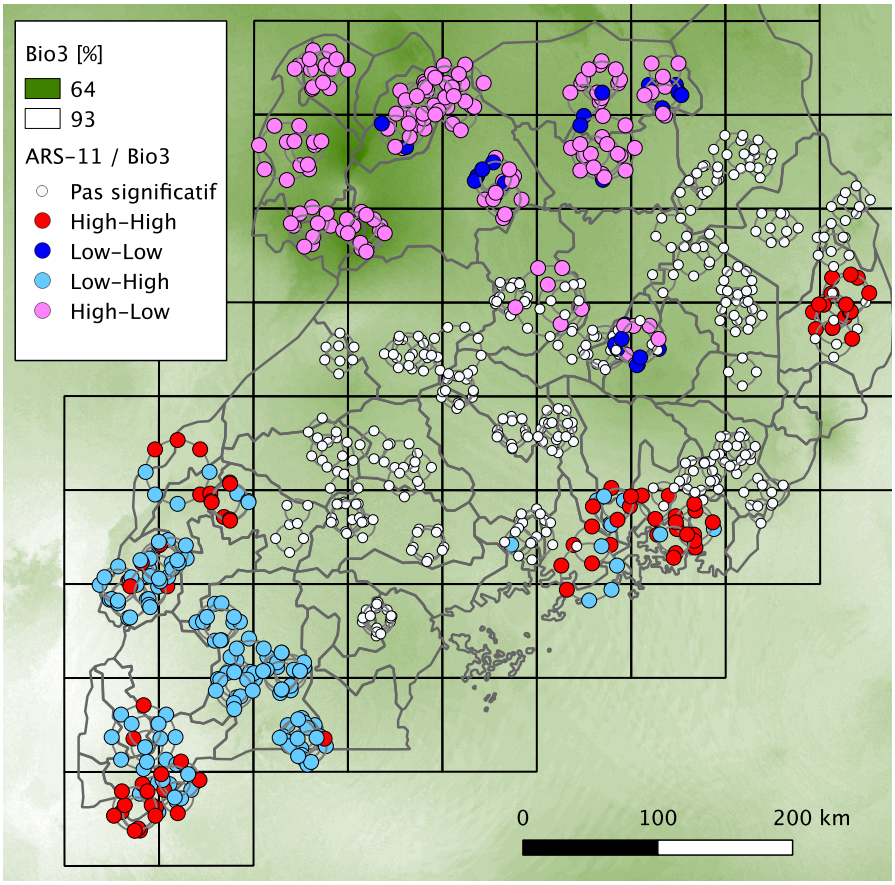
Autocorrélation spatiale locale bivariée

Les LISA bivariés mesurent la corrélation spatiale locale entre un marqueur et une variable environnementale. J'ai calculé les LISA bivariés correpondant aux trois marqueurs pour deux variables environnementales avec le logiciel `GeoDa` [14]. La première série de cartes inclut la covariable isothermalité « bio3 » (figure 7.20) et la seconde série inclut le coefficient d'appartenance à la population ankole, « ankole » (figure 7.21) .

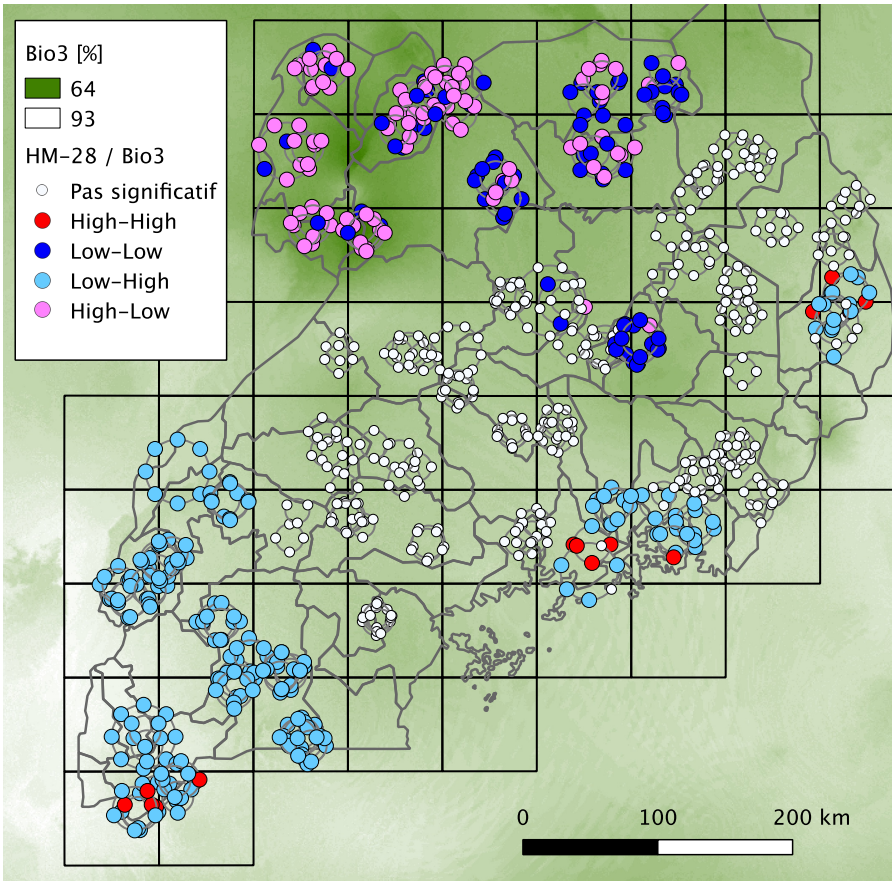
Les associations entre les trois marqueurs et « bio3 » sont significatives sur environ la moitié du territoire, mais elles ne sont pas significatives au centre du pays ainsi que dans une région située dans l'est. Les marqueurs ARS-11 et HM-28 sont fréquents dans le nord et l'est de l'Ouganda, alors que « bio3 » est plus basse au nord qu'au sud. Les associations dans le sud et l'est sont donc principalement de type « low-high » (avec quelques « high-high ») alors qu'elles sont de type « high-low » et « low-low » dans le nord (indices globaux $I = -0,39$ et $I = -0,44$). A l'inverse, le marqueur ARS-94 est principalement présent dans le sud-ouest du pays, où la valeur de l'isothermalité est élevée, c'est pourquoi les corrélations locales de ce marqueur sont généralement positives (de type « high-high » et « low-low », indice global $I = 0,22$). La principale différence entre HM-28 (qui est détecté à l'unanimité) et les deux autres marqueurs est qu'il est moins fréquent dans l'est et présente donc des associations spatiales plus homogènes, « low-high » dans le sud-ouest et l'est ainsi qu'un gradient de « low-low » à « high-low » du centre au nord-ouest du pays.

Les trois marqueurs présentent une association significative avec la variable « ankole » sur la plus grande partie du territoire (fig. 7.21). Les marqueurs ARS-11 et HM-28 sont corrélés négativement à cette variable car ils apparaissent principalement dans le nord du pays où la valeur de « ankole » est faible (indices globaux $I = -0,55$ et $I = -0,35$). Les associations sont donc souvent « low-high » ou « high-low ». Ces marqueurs se distinguent principalement par la présence plus fréquence du premier dans l'est et par ses quelques occurrences dans le sud-ouest du pays ; ARS-11 présente donc plus de points « high-high » dans le sud-ouest et plus de points « high-low » dans l'est du pays que HM-28. Le marqueur HM-28 est d'ailleurs pratiquement absent dans tout le sud du pays. Comme son association avec « ankole » est significative dans le centre du pays, elle présente clairement un cluster « low-high » dans le sud et le centre et un gradient « low-low » - « high-low » de l'est au nord-ouest du pays. Comme le marqueur ARS-94 est principalement présent dans le sud de l'Ouganda, son association à « ankole » est généralement de type « high-high » ou « low-low » (indice global $I = 0,36$.

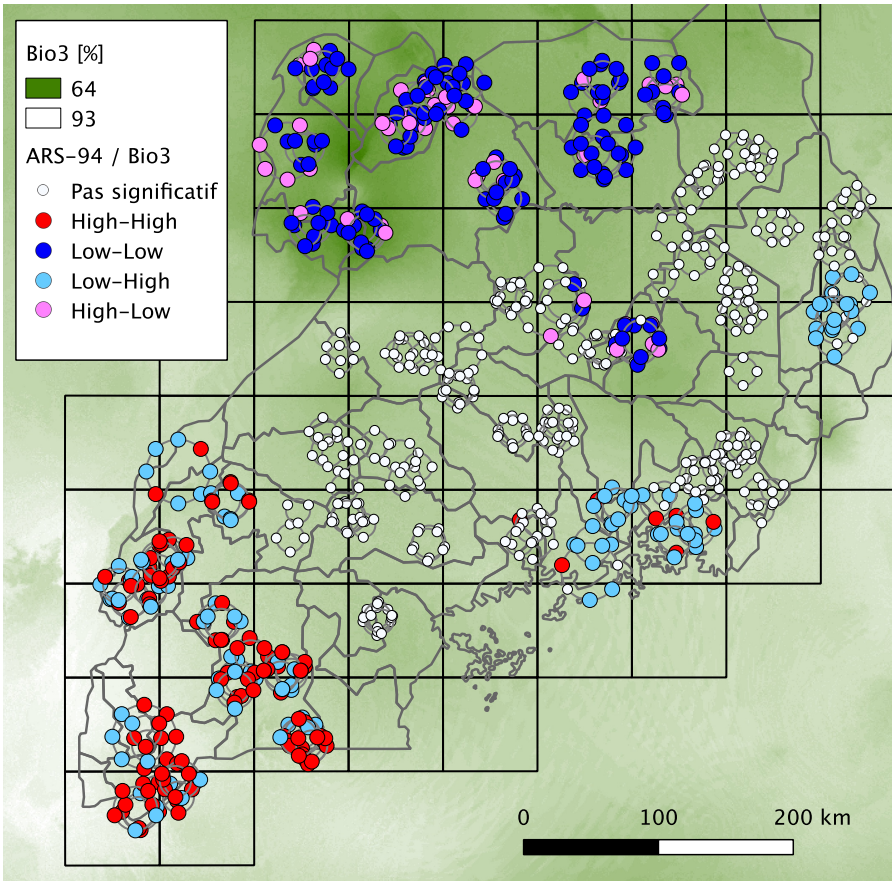
14. http://geodacenter.asu.edu/



(a) ARS-11

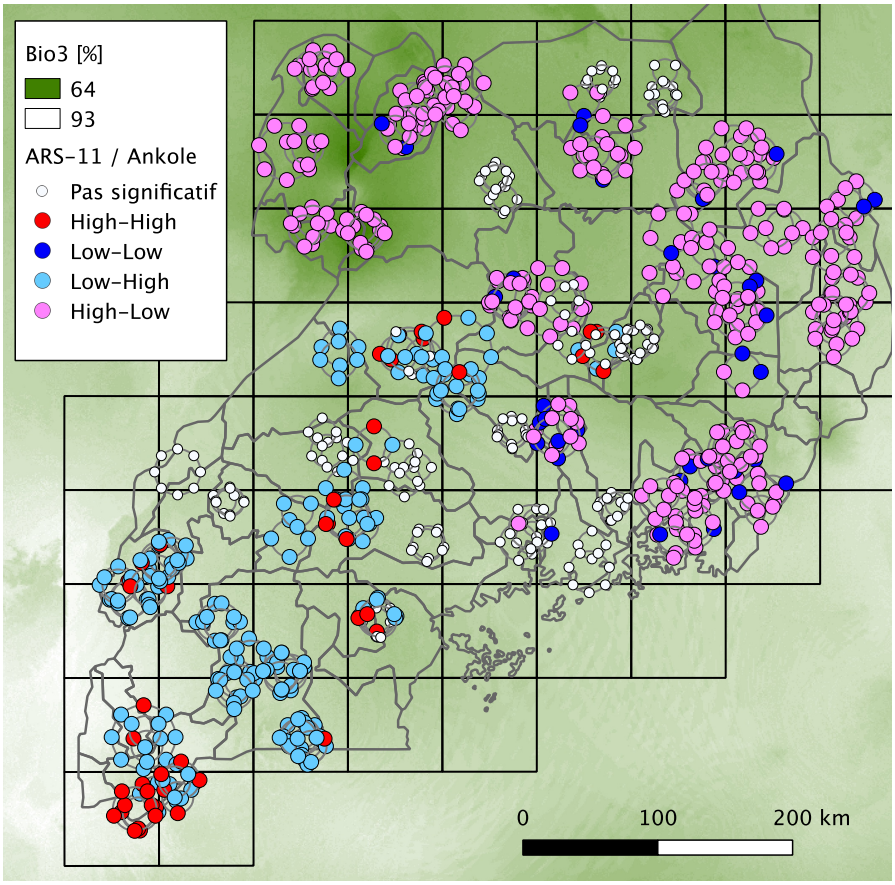


(b) HM-28

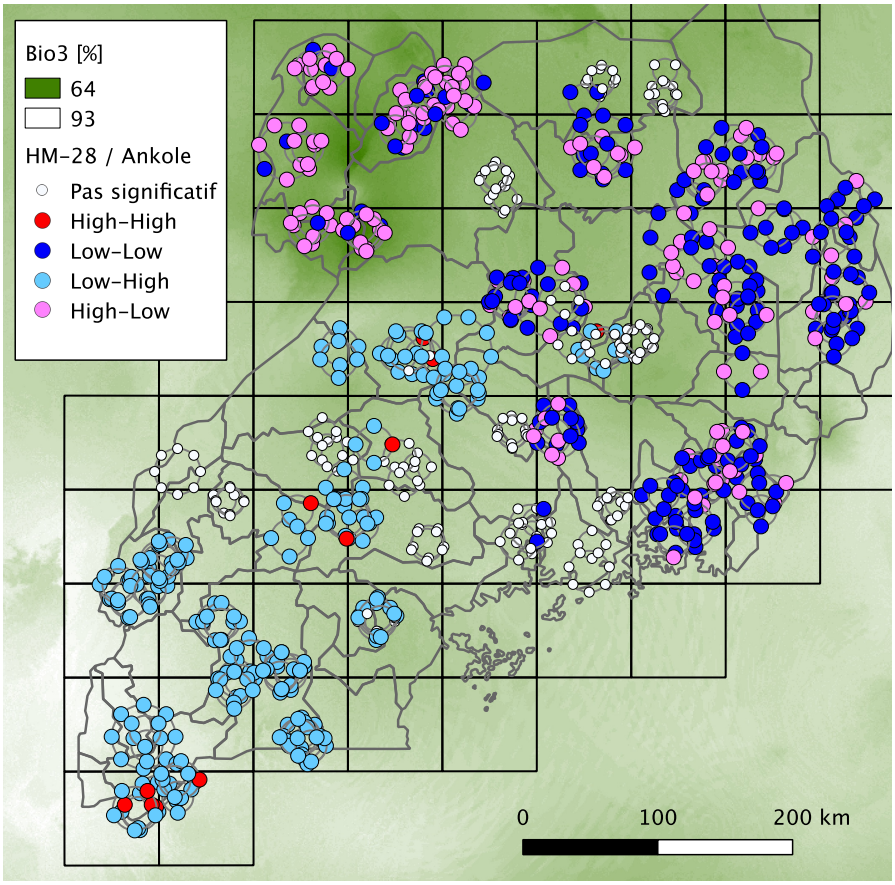


(c) ARS-94

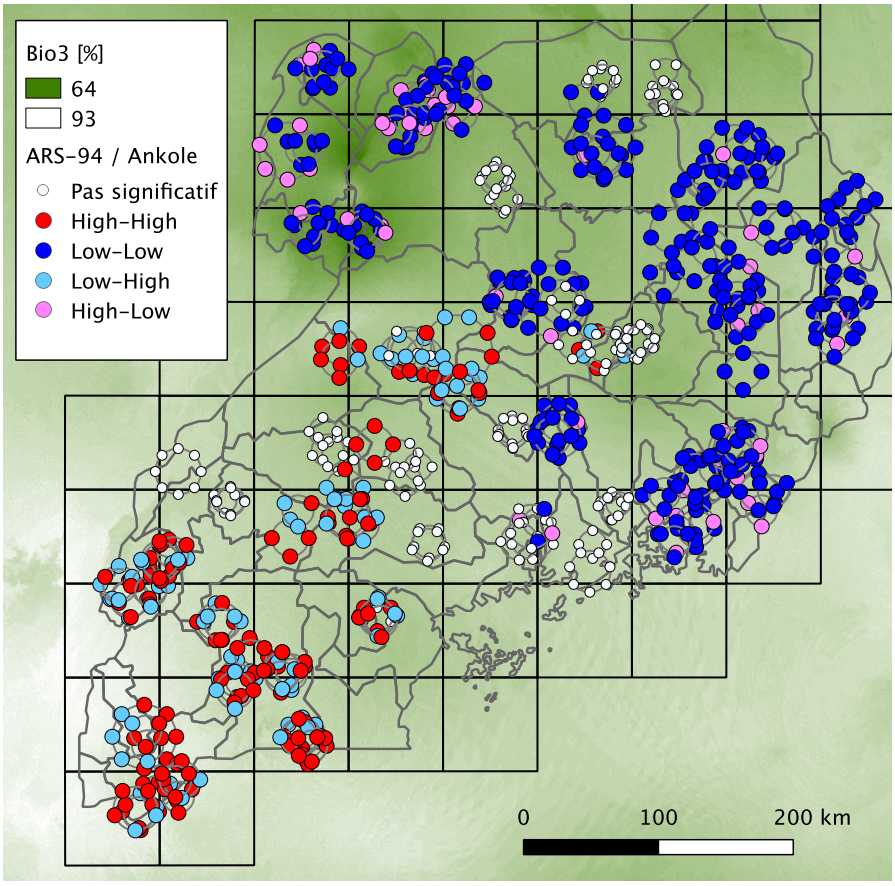
Figure 7.20 – Carte des indices d'autocorrélation spatiale locale bivariée des marqueurs ARS-11, HM-28 et ARS-94 issus des données 54k (LISA bivariés). La covariable est l'isothermalité (cf table 4.2 et corps du texte). La pondération est basée sur les 20 plus proches voisins. Les points rouges indiquent les marqueurs présents dans des lieux où la variable environnementale a une valeur élevée (« *high-high* ») ; si la variable a une valeur basse, le point est rose (« *high-low* »). Les points bleu marine indiquent les marqueurs absents dans des lieux où la variable a une valeur basse (« *low-low* ») ; si la variable a une valeur élevée, le point est bleu ciel (« *low-high* »). Les point blancs signalent les valeurs non significatives ($p > 0.001$). Le fond de carte représente l'isothermalité (cf table 4.2 et corps du texte).



(a) ARS-11



(b) HM-28



(c) ARS-94

Figure 7.21 – Carte des indices d'autocorrélation spatiale locale bivariée des marqueurs ARS-11, HM-28 et ARS-94 issus des données 54k (LISA bivariés). La covariable est le coefficient d'appartenance à la population ankole (voir fig. 7.2b). La pondération est basée sur les 20 plus proches voisins. Les points rouges indiquent les marqueurs présents dans des lieux où la variable environnementale a une valeur élevée (« *high-high* ») ; si la variable a une valeur basse, le point est rose (« *high-low* »). Les points bleu marine indiquent les marqueurs absents dans des lieux où la variable a une valeur basse (« *low-low* ») ; si la variable a une valeur élevée, le point est bleu ciel (« *low-high* »). Les point blancs signalent les valeurs non significatives ($p > 0.001$). Le fond de carte représente l'isothermalité (cf table 4.2 et corps du texte).

7.2 Validation avec données simulées

Les comparaisons entre approches de détection de signatures de sélection sont confrontées à l'incertitude concernant l'influence réelle de la sélection sur les marqueurs analysés. L'utilisation de données simulées permet de tester le comportement d'une méthode dans un cadre où les marqueurs neutres et adaptatifs sont clairement identifiés. Afin de mieux comprendre leur détections respectives, j'ai comparé les résultats de Samβada et `LFMM` utilisés lors de l'analyse d'un jeu de 100 SNPs simulés avec `CDPOP`. J'ai utilisé une population de 5'000 individus soumis à une pression de sélection dépendant de la latitude (cf sec. 4.6 et Jones et al., 2013). J'ai extrait un sous-ensemble des individus pour obtenir un jeu similaire aux données 800k. Pour chaque simulation, j'ai découpé le territoire selon une grille de 7x7 cellules et j'ai sélectionné deux individus aléatoirement dans chaque cellule, pour imiter le choix des bovins pour le génotypage à 800k SNPs. L'échantillon ainsi formé comporte 98 individus. J'ai recodé les données en 300 marqueurs binaires (correspondant aux génotypes AA, AG et GG respectivement) pour Samβada et 100 marqueurs comptant les allèles "A" pour `LFMM` (AA=2, AG=1, GG=0). Pour ce dernier, j'ai déterminé le nombre de populations avec `Admixture`. J'ai analysé une simulation par intensité de sélection, il s'est avéré que la meilleure partition était donnée par $K = 2$.

La table 7.33 présente les résultats obtenus avec Samβada et `LFMM`. Le seuil de significativité était fixé à $\alpha = 0.01$ avant la correction de Bonferroni. Les résultats de Samβada sont basés sur le score G. Le fait important est que les deux approches n'ont pas identifié les loci soumis à la sélection pour l'intensité faible, et que Samβada a détecté trois faux positifs. Samβada et `LFMM` ont détecté les trois mêmes loci adaptatifs pour l'intensité moyenne, et tous les loci soumis à la sélection pour l'intensité forte. Ils n'ont pas trouvé de faux positifs dans ces cas. Lorsque Samβada a détecté le locus adaptatif, il a identifié l'allèle GG comme étant soumis à la sélection, parfois également l'allèle AA.

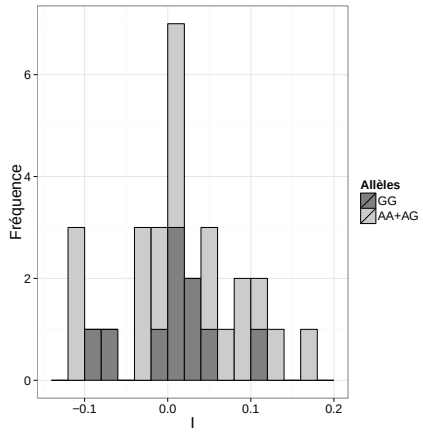
		Intensité de sélection		
		Faible (1%)	Moyenne (10%)	Forte (50%)
Loci sous sélection détectés	Samβada	0	3	10
	`LFMM`	0	3	10
Fausses découvertes	Samβada	3	0	0
	`LFMM`	0	0	0

Table 7.33 – Résultats de Samβada et de `LFMM` pour les donnés simulées. Pour chaque intensité de sélection, la mortalité infantile maximale est indiquée entre parenthèses.

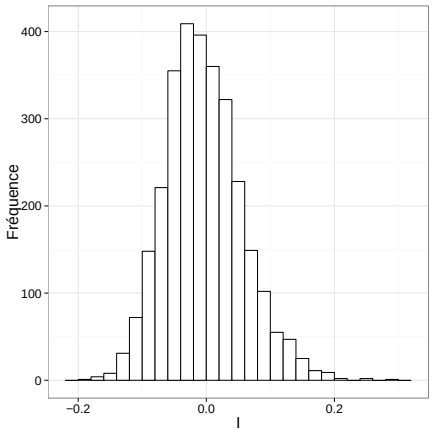
Samβada a également mesuré l'autocorrélation spatiale en utilisant les 5, 10, 15, 20, 25, 40 et 50 plus proches voisins. La figure 7.22 présente la distribution des I de Moran globaux basés sur 5 voisins (soit $\sim 5\%$ des points). La distribution des marqueurs neutres (colonne de droite : b, d et f) ne dépend pas de l'intensité de la sélection. La colonne de gauche présente les marqueurs binaires correspondant au locus soumis à la sélection. La distribution pour l'intensité faible est similaire à celle des loci neutres. Dans les deux autres scénarios, l'autocorrélation de certains

marqueurs augmente. Dans le cas de l'intensité forte, les marqueurs GG sont clairement plus autocorrélés que les marqueurs neutres. Ces résultats seront discutés à la sec. 8.2.7.

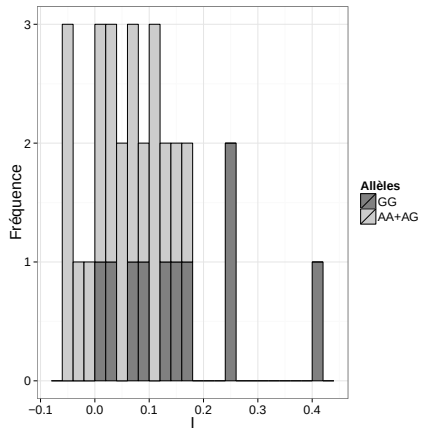
Ceci clôt la présentation des résultats. Le chapitre suivant est consacré à la discussion.



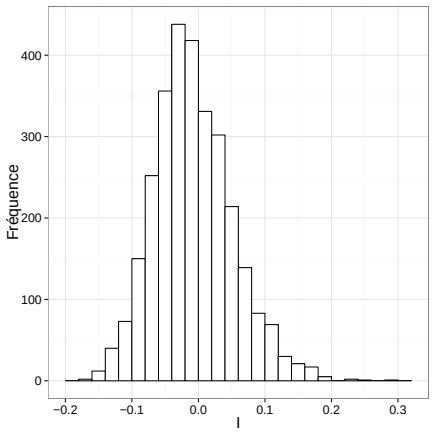
(a) Intensité faible, loci adaptatifs



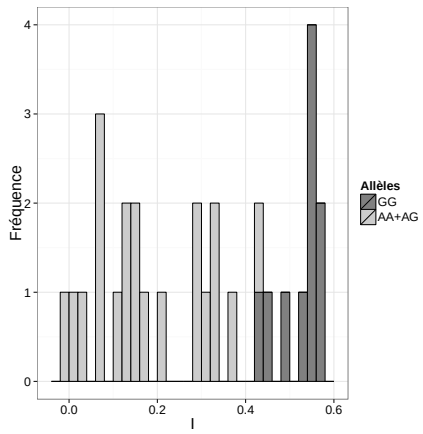
(b) Intensité faible, loci neutres



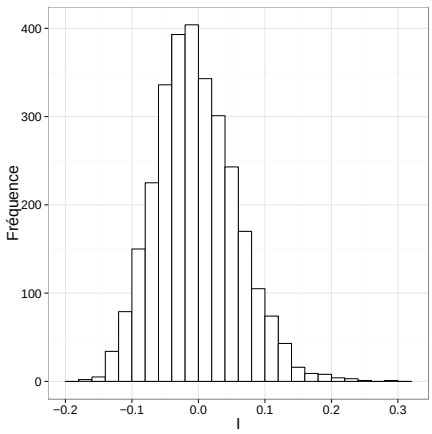
(c) Intensité moyenne, loci adaptatifs



(d) Intensité moyenne, loci neutres



(e) Intensité forte, loci adaptatifs



(f) Intensité forte, loci neutres

Figure 7.22 – Distribution des *I* de Moran globaux pour les données simulées. Les simulations ont été regroupées par intensité de sélection : faible (1% de mortalité, a et b), moyenne (10% de mortalité, c et d) et forte (50% de mortalité, e et f). La colonne de gauche présente les loci adaptatifs et celle de droite les loci neutres. Chaque intensité de sélection a fait l'objet de 10 simulations indépendantes, avec à chaque fois un locus adaptatif et 99 loci neutres. Les marqueurs génétiques sont des SNPs bialléliques qui ont été recodés en trois variables binaires par locus.

8 Discussion

Dans cette partie, je vais d'abord présenter quelques remarques sur la collecte des données. Puis, sur la base des résultats du chapitre précédent, je vais discuter des avantages et des inconvénients des approches corrélatives ainsi que de l'apport des statistiques spatiales avant de comparer les méthodes utilisées. J'aborderai ensuite la prise en compte de la structure de population dans les analyses de Samβada et finirai avec quelques observations sur sa diffusion.

8.1 Collecte des données

La fiabilité des résultats obtenus lors d'une étude en écologie dépend essentiellement de la stratégie employée pour collecter les données (Albert et al., 2010). Une étude en génomique environnementale requiert de caractériser les habitats des organismes étudiés et d'échantillonner ces derniers dans leurs différents environnements possibles.

8.1.1 Précautions pour le choix des données environnementales

Les approches corrélatives utilisent des variables environnementales pour décrire l'habitat des organismes étudiés. L'étape préparatoire consistant à réunir ces données devrait tenir compte de la taille de la région étudiée afin de choisir quelle résolution spatiale utiliser. Les bases de données disponibles en ligne fournissent des données topo-climatiques à haute résolution. Celles-ci permettent de décrire précisément chaque lieu où des échantillons ont été récoltés. Or certains phénomènes ne sont détectables qu'à une échelle appropriée, qui n'est pas forcément celle de la plus haute résolution des données. Dans ce cas, une analyse utilisant des données multi-échelles est susceptible de détecter ces relations (Leempoel et al., 2013). Les études concernant les animaux devraient peut-être également tenir compte de ce phénomène, car leur habitat est probablement plus varié que le lieu de l'échantillonnage.

Les variables environnementales doivent être aussi variées que possible pour représenter un maximum de pressions de sélection potentielles. Cet ensemble peut être complété avec

des « variables inconnues », comme les MEMs, qui modélisent l'effet des conditions environnementales non-mesurées (Manel et al., 2010). Lors de l'étude de grands jeux de données génétiques, il faut prendre en compte le temps nécessaire pour traiter tous les modèles, car celui-ci croît linéairement avec le nombre de variables environnementales.

Beaucoup d'études emploient des modèles multivariés pour décrire l'influence combinée de plusieurs facteurs environnementaux. La pertinence de l'utilisation de nombreuses variables environnementales doit alors être soigneusement évaluée. D'une part, le temps de calcul croît avec le carré du nombre de variables (dans le cas bivarié). D'autre part, les corrélations entre prédicteurs augmentent la variance des paramètres estimés et peuvent produire des modèles instables dont les coefficients varient considérablement selon le nombre de points analysés. La multicollinéarité des variables environnementales doit être mesurée pendant la préparation des données. Un sous-ensemble des variables peut être sélectionné pour limiter cette inflation de la variance.

8.1.2 Récolte des échantillons

Dans une étude où les populations seraient clairement séparées, la divergence entre ces populations pourrait être modélisée par une variable caractéristique d'appartenance. Toutefois, même des populations géographiquement distinctes peuvent présenter des corrélations de fréquences alléliques (Coop et al., 2010). L'analyse pourrait également être conduite séparément sur chaque population pour déterminer quels loci sont détectés dans les différentes régions. Cette approche s'inspire des études démontrant que l'adaptation locale à la même pression de sélection peut influencer des régions différentes du génome pour des populations distinctes.

La planification d'une étude en génomique environnementale devrait viser à diversifier les lieux et habitats échantillonnés plutôt que de se baser sur le nombre d'individus analysés par population. Les caractéristiques du terrain et des habitats présents dans la zone d'étude peuvent être déterminées avant la campagne en analysant les variables environnementales déjà disponibles. Une stratégie d'échantillonnage consistant à récolter des échantillons dans plusieurs régions différentes pour chaque type d'habitat permettrait certainement de séparer les effets démographiques de ceux liés à la sélection (Schwartz et McKelvey, 2009 ; Manel et al., 2012).

Dans le cadre d'une campagne à large échelle, la collecte des échantillons pourrait également être facilitée en utilisant des *smartphones* ou des tablettes tactiles. Ces appareils intègrent aujourd'hui un capteur GPS et un appareil photo d'assez bonne qualité. En développant une application dédiée durant la préparation de la campagne, les échantillonneurs pourraient relever la position des individus, les photographier et enregistrer les données morphologiques en utilisant un seul support. La mise en commun des données sur un serveur central s'en trouverait aussi simplifiée et diminuerait le risque de commettre des erreurs de saisie.

8.2 Détection de la sélection naturelle

Lors d'une étude de l'adaptation locale, l'échantillonnage est suivi de l'extraction de l'information génétique. Puis vient la détection des signatures de sélection naturelle, cette étape est discutée ici sur la base des résultats obtenus au chapitre précédent.

Le cheptel bovin ougandais se décline en deux populations, les vaches ankoles au sud-est et les zébus à cornes courtes au nord-est, qui se mélangent au centre du pays. Les analyses de structure des populations concordent et séparent la population bovine ougandaise en deux groupes principaux. Certains des marqueurs génétiques identifiés comme étant potentiellement soumis à la sélection naturelle portent donc la marque de cette structure de population.

8.2.1 GLM et faux positifs

Effet démographique

Samβada détecte beaucoup de loci potentiellement soumis à la sélection dans les données 54k (2'500 SNPs sur 40'034 soit 6,2% pour un seuil $\alpha = 0,01$). Comme la répartition spatiale des zébus et des ankoles se superpose à certaines variables environnementales présentant un gradient nord-sud, les loci détectés ne sont probablement pas tous soumis à la sélection, et certains doivent probablement dépendre de la structure de population. Une telle structure peut par exemple se former si une population de taille finie se retrouve séparée en deux sous-populations isolées. L'apparition de nouvelles mutations et la dérive génétique (due au tirage aléatoire des gamètes dans une population finie) modifiera graduellement le patrimoine génétique des deux populations. La mesure d'un indice de différentiation comme F_{ST} montrera alors que les populations sont génétiquement distinctes [1]. Si les habitats de ces deux populations sont différents — et il est d'ailleurs fort probable qu'au moins une variable environnementale permette de les distinguer — une approche corrélative identifiera les marqueurs génétiques qui ont divergé et attribuera cette différentiation à la variable environnementale, bien que cette divergence n'ait rien à voir avec la sélection naturelle. C'est pourquoi les approches corrélatives ont tendance à détecter de fausses signatures de sélection naturelle dans les populations structurées. La situation en Ouganda est celle de deux populations qui se rencontrent après une longue séparation — la divergence de *B. taurus* et *B. indicus* remonte à au moins 330'000 ans, bien avant leur domestication (Ajmone Marsan et al., 2010) — et les différences génétiques dues à leur histoire démographique produisent des faux positifs.

De manière à aborder l'analyse des effets démographiques avec Samβada, les animaux ont été assignés à une population en fonction de la fraction majoritaire de leur génome : certains ankoles portent des gènes d'origine zébu et quasiment tous les zébus ont une ascendance

1. Le même phénomène peut arriver si les deux populations restent connectées mais que le nombre d'individus migrant à chaque génération ne suffit pas à compenser les mutations et la dérive génétique.

ankole (cf fig. 7.5, p. 100). Seule une petite partie des loci détectés dans les données 54k sont aussi détectés dans les deux populations individuellement. Un locus est détecté par le test *G* chez les ankoles et aucun avec le test de Wald, alors que 38 loci sont détectés chez les zébus avec *G* et 5 avec Wald. Pourtant chaque population est représentée par trois ou quatre cents individus, ce qui devrait être suffisant pour cette analyse. Cela signifie peut-être que la plupart des loci identifiés sur la base de tous les individus discriminent les populations. De même, la différence entre le nombre de détections chez les zébus et les ankoles pourrait provenir des deux sous-populations de zébus identifiées lors de la classification en sept clusters par `Admixture` (cf fig. 7.5b p. 100 pour la structure de population et fig. 7.6 p. 7.6 pour la carte). Ces sous-populations pourraient créer un facteur de confusion d'une intensité moindre que celui dû aux populations zébu et ankole. A ce point, il convient de remarquer que la longitude figure parmi les variables explicatives dans les modèles significatifs chez les zébus et qu'il pourrait s'agir d'un facteur de confusion lié à la distribution de ces deux sous-populations. Leur présence forme un gradient est-ouest dans le Nord de l'Ouganda et cette distribution spatiale est corrélée à la longitude (cf table 7.11a p. 115 et fig. 7.6 p. 101). S'il y a réellement une différence entre ces deux groupes de zébus, elle pourrait découler de différences dans les méthodes d'élevage suivant les régions d'Ouganda. Les deux sous-populations de zébus observées dans la classification en sept populations pourraient être également un artefact dû à un nombre de populations attendues supérieur à la structure réelle. Dans ce cas, `Admixture` pourrait résoudre la contradiction en découpant un cluster en deux et en divisant les coefficients d'appartenance à ce cluster en deux partie égales pour les nouveaux clusters (cf la coupure horizontale entre les population 5 et 6 sur la fig. 7.5b p. 100). Cela expliquerait pourquoi les performances des classifications en quatre et sept populations sont équivalentes sur la fig. 7.1 p. 95.

Maintenant, ces détections chez les zébus s'expliquent peut-être par le choix des animaux lors du développement de la puce à ADN. Plusieurs races de zébus ont alors été analysées et intégrées, et il est possible que les régions variables de leur génome soient mieux représentées que celles du génome des ankoles.

La faible proportion de loci détectés avec l'aide de ces modèles démographiques peut également indiquer que certains SNPs sont fixés dans une des populations et que ce balayage sélectif complet (*hard selective sweep*) n'est pas observable sans comparaison avec une autre population. Cela expliquerait pourquoi les ankoles, qui sont peu hybridés avec d'autres races, ne présentent que quelques loci détectés. A l'inverse, les loci détectés chez les zébus reflèteraient la fraction du patrimoine génétique d'origine ankole dans la population : des SNPs qui seraient fixés chez les ankoles seraient alors uniquement détectés chez les zébus. Il reste à déterminer si cette hypothèse peut être testée sur la base de nos données.

D'un autre point de vue, le peu de loci détectés provient peut-être de la faible variabilité des conditions environnementales. Le climat ougandais étant assez homogène et les deux populations de vaches vivant dans des régions globalement disjointes, la variabilité des conditions environnementales est peut-être trop faible dans ces zones pour observer une corrélation

nette avec la présence d'allèles sous sélection.

Finalement, le faible nombre de détections dans les populations zébu et ankole pourraient provenir de la diminution du nombre d'individus qui baisserait la puissance des tests statistiques. L'application de la correction de Bonferroni avec $\alpha = 0,01$ provoque peut-être des faux négatifs.

Nombre de marqueurs génétiques élevé

D'un autre point de vue, le nombre de marqueurs détectés avec tous les individus pourrait être élevé à cause du nombre important de marqueurs analysés. En effet, la proportion de marqueurs détectés (6%) est comparable à celle d'études antérieures (p. ex. Poncet et al., 2010 ; Manel et al., 2010). Comme ces études utilisent aussi des approches corrélatives, leurs résultats comportent très probablement des fausses découvertes [2]. Par conséquent, tous les loci identifiés par Samβada ne sont pas forcément adaptatifs, mais leur proportion correspond aux chiffres produits par des études déjà menées en génomique environnementale. Notons également que les études susmentionnées utilisent des marqueurs AFLPs qui sont généralement répartis sur tout le génome et ne sont pas sujets au biais de détermination (*ascertainment biais*) (Schwartz et al., 2009). Par contre, les vaches ougandaises ont été génotypées à l'aide d'une puce à ADN standard, principalement destinée à la sélection des animaux d'élevage. Les SNPs utilisés ont donc déjà été identifiés et choisis lors du développement de la puce. Certains de ces loci ont été inclus pour leur capacité à distinguer les populations ou à analyser la qualité du lait et de la viande (Illumina Inc., 2012a ; Illumina Inc., 2012b). De ce fait pourrait résulter un facteur de confusion avec l'environnement. De plus, le choix d'un ensemble de SNPs à inclure sur une puce peut mener à un biais de détermination : les marqueurs retenus sont représentatifs de la variabilité présente dans le groupe d'individus séquencés. Les individus genotypés avec la puce possèdent probablement d'autres mutations et les études menées avec des SNPs risquent donc de sous-estimer la variabilité génétique de races pour lesquelles la puce n'a pas été prévue. Toutefois, environ 10'000 SNPs considérés ici ont été identifiés par le projet Bovine HapMap (Illumina Inc., 2012a ; Illumina Inc., 2012b) et certains d'entre eux doivent avoir une valeur adaptative (The Bovine Genome Sequencing and Analysis Consortium et al., 2009).

Qualité et nombre d'individus

Le nombre de loci détectés dans les données 54k est également lié au nombre d'individus considérés. Plusieurs individus ont été échantillonnés dans chaque ferme et, malgré les précautions prises, certains sont peut-être apparentés. Les données peuvent donc contenir des pseudo-réplicats ce qui implique que certains marqueurs sont autocorrélés. La taille de l'échantillon est comparable à celle d'études antérieures, qui comprenaient moins de

2. Poncet et al. utilisent deux populations indépendantes pour restreindre la liste des marqueurs candidats, mais la proportion de marqueurs détectés dans chaque population est de l'ordre de 5-10%.

marqueurs (Poncet et al., 2010 ; Manel et al., 2010). Une analyse portant sur 800 individus est suffisamment puissante pour détecter des signatures de sélection dans un jeu de 40k SNPs avec la correction de Bonferroni.

A l'inverse, le fait qu'aucun modèle ne passe le test de Wald parmi les données 800k avec la correction de Bonferroni est certainement dû à la petite taille de l'échantillon. Le nombre de modèles à considérer pour le score de Wald se prête à une discussion informelle. La distance moyenne entre deux SNPs inclus sur la puce est de 3,4 kbp ce qui est relativement proche. Certains marqueurs sont donc en déséquilibre de liaison et ne sont par conséquent pas indépendants. De plus, le recodage des SNPs, qui est nécessaire à Sambada, fait que la somme des trois marqueurs binaires correspondant au même SNP est toujours égale à 1 (car un individu possède exactement un génotype pour chaque SNP), ce qui fait qu'un tiers des marqueurs sont redondants. Nous pourrions ainsi considérer un « nombre effectif » de marqueurs qui serait inférieur au nombre de génotypes utilisés. Nous pourrions aussi ne considérer qu'une seule variable environnementale, ce qui contribuerait à diminuer drastiquement le « nombre effectif » de modèles analysés. Outre son manque apparent de rigueur, cette approche ne permet pas de baisser le seuil de significativité calculé avec la correction de Bonferroni suffisamment bas pour que les modèles ayant les plus hauts scores de Wald deviennent significatifs. C'est pourquoi la détection des signatures de sélection pour les données 800k est basée sur le test du rapport de vraisemblance avec la correction de Bonferroni. De la sorte, les 57 SNPs identifiés avec le seuil $\alpha = 0,01$ représentent 0,01% du total (cf 7.8 p. 112). Certains de ces loci sont regroupés dans les mêmes régions du génome (cf fig. 7.8, p. 107) et l'analyse des fonctions des gènes auxquels ils sont potentiellement liés serait donc plus aisée qu'avec les données 54k, où les loci sont répartis sur le génome (cf fig. 7.7, p. 105). Notons également qu'une sélection des modèles basée sur le taux de fausses découvertes selon Storey et Tibshirani permettrait d'utiliser le score de Wald (cf tab. 7.9, p. 113). Avant d'appliquer cette approche FDR, il faudrait vérifier si elle est compatible avec la sélection des modèles multivariés basées sur les parents, ce que je n'ai pas encore fait. C'est pourquoi j'ai utilisé la correction de Bonferroni avec le score G pour l'analyse des données 800k, afin que les tests employés soient cohérents entre les cas uni- et multivariés.

Si l'on considère maintenant les résultats obtenus avec les données 800ksub, ils peuvent paraître surprenants à première vue. L'analyse porte sur le même nombre d'individus que les données 800k, mais avec moins de marqueurs : le seuil de significativité pour le score G baisse, alors que les modèles associés aux loci considérés, et donc leurs scores, sont les mêmes que précédemment. Nous pourrions nous attendre à ce que le nombre de modèles détectés augmente, puisque le seuil est plus bas. Or, avec $\alpha = 0,01$, seuls 12 modèles et 7 loci, soit 0,02% du total, sont détectés. La proportion des loci potentiellement soumis à la sélection naturelle donne des chiffres similaires aux données 800k. Les marqueurs composant ces jeux de données sont répartis relativement uniformément sur le génome, par conséquent la fraction des SNPs situés dans les régions potentiellement soumises à la sélection naturelle est sensiblement égale entre les données 800k et 800ksub. Les loci détectés se regroupent en quatre clusters dans chaque jeu de données et deux de ces clusters se situent dans les

mêmes régions génomiques pour les deux jeux de données. L'utilisation d'un sous-ensemble de marqueurs ne fournit donc pas exactement les mêmes résultats que le jeu complet. Les marqueurs identifiés avec le sous-ensemble de SNPs sont probablement plus éloignés du locus sous sélection puisque l'échantillon de loci est moins dense.

La comparaison des données 800ksub et 54k montre que, malgré un nombre de marqueurs équivalent (30k et 40k), les résultats sont très différents. Les 7 loci détectés dans les données 800ksub sont groupés en quatre clusters alors les 2'500 loci détectés dans les données 54k sont répartis sur tout le génome. Les analyses présentent deux différences principales. Premièrement, le nombre d'individus est divisé par huit dans le jeu 800ksub, ce qui diminue la puissance du test et certains modèles ne sont plus désignés comme significatifs. En particulier, aucun modèle n'est détecté avec le test de Wald. Deuxièmement, il convient de souligner que les 102 individus génotypés pour les données 800k et 800ksub proviennent de fermes différentes. Ces jeux de données ne contiennent donc pas de pseudo-réplicats, mais certains marqueurs pourraient quand même être autocorrélés. Ce nombre de marqueur est également différent : environ 5'000 SNPs de la puce 54k ont été écartés de la puce 800k. Ils ont peut-être été considérés comme peu informatifs ou redondants. Ceci n'explique cependant pas la différence au niveau du nombre de détections puisque 2'247 des 2'500 SNPs identifiés dans les données 54k (soit 90%) se retrouvent dans les données 800ksub. Quelle est alors l'origine de cette différence ? Est-ce un manque de puissance de détection pour 102 individus ? Ou est-ce un effet de la structure de population, ou des pseudo-réplicats, dans les données 54k ?

Efficacité de Samβada

L'analyse des modèles univariés montre que l'approche corrélative de Samβada permet de détecter des loci potentiellement soumis à la sélection naturelle dans de grands jeux de données moléculaires. Le traitement des données 800k a duré moins de trois heures pour calculer 43 millions de modèles sur un ordinateur de bureau. Cela veut dire par exemple que pour les données sur les chèvres et les moutons au Maroc pour lesquelles le consortium NextGen attend environ 16 millions de SNPs pour 160 individus, les calculs devraient durer une centaine d'heures à répartir sur plusieurs machines. En ce qui concerne la distribution des p-valeurs, la FDR selon Storey et Tibshirani est bien adaptée à Samβada et permettrait de filtrer les modèles avec un taux de faux positifs prédéfini. En revanche, Samβada ne peut pas déterminer dans le cas présent quels marqueurs constituent de fausses découvertes dues à la structure de population (effet démographique). En particulier, les loci présentant les plus hauts scores G et Wald seront les premiers détectés quelle que soit la correction utilisée pour les comparaisons multiples. Cependant, l'analyse spatiale des marqueurs détectés, notamment au moyen du I de Moran, pourrait permettre de discriminer les fausses découvertes.

Modèles multivariés

L'étude des modèles multivariés poursuit deux objectifs. Le premier, dans la foulée de la section précédente, est de déterminer si l'inclusion dans le modèle de paramètres décrivant la structure de population permet de discriminer entre les loci soumis à la sélection et les découvertes fortuites. Le deuxième est d'analyser si une combinaison de prédicteurs permet une meilleure description de la distribution d'un allèle qu'un modèle univarié.

Les modèles bivariés peuvent permettre d'affiner les prédictions de présence des allèles, à moins qu'une variable n'explique une grande partie de la distribution des marqueurs. Par exemple, la table 7.12a montre que la variable « ankole » est souvent le meilleur prédicteur pour les modèles univariés. Les 4'354 SNPs détectés avec « ankole » sont probablement liés d'une manière ou d'une autre à la structure de population et la présence de nombreux allèles peut donc être prédite à partir de cette structure. Cependant certains SNPs soumis à la sélection sont peut-être inclus dans ce groupe, notamment ceux susceptibles d'être fixés dans une population. Les 398 SNPs non détectés avec « ankole » mais détectés avec une autre variable sont peut-être sous sélection indépendamment de la population. Certains SNPs sont détectés avec un modèle bivarié incluant la variable « ankole ». Ces marqueurs sont donc significativement corrélés à la deuxième variable explicative alors que le modèle tient compte de la structure de population. Trois de ces marqueurs sont identifiés avec un modèle bivarié alors que leur modèle parent incluant la variable « ankole » est aussi significatif. Cette sélection de modèles est très conservatrice mais produit des résultats cohérents avec les autres méthodes. Ces modèles bivariés incluent une variable pour la structure de population et une variable environnementale, ce qui rappelle le modèle utilisé par `LFMM`.

La variable latitude est souvent présente dans les modèles univariés de la table 7.4 (p. 109). Ce sont d'ailleurs ces modèles qui ont suggéré le possible facteur de confusion de la structure de population dans les résultats. La latitude se comporte comme une première approximation de la la variable « ankole ».

Concernant le deuxième point maintenant, la table 7.12b (p. 118) présente des modèles pour lesquels une combinaison additive de prédicteurs fournit une meilleure prédiction de la présence d'un allèle que chacune de ces variables prise séparément. Cependant il faut également comparer l'ensemble des modèles uni- et bivariés obtenus pour un marqueur. Il peut arriver qu'une variable environnementale fournisse une meilleure prédiction qu'une combinaison de deux autres variables. Le score AIC, basé sur la vraisemblance, permet justement de faire cette comparaison : le plus petit AIC indique quel modèle est le plus efficace. La table 7.12a montre que, pour certains marqueurs, les modèles univariés impliquant la variable « ankole » fournissent une meilleure estimation de la probabilité de présence de l'allèle que les modèles bivariés présentés. Cet effet est accentué par la présentation de la table 7.12b : les modèles y sont classés selon leur score G par analogie avec le cas univarié, pour lequel les scores G et AIC sont directement liés. Cet ordre favorise les modèles ayant des parents avec une vraisemblance relativement basse, mais suffisamment haute pour être significative. A l'inverse,

les modèles bivariés significatifs ayant des parents avec un score *G* élevé auront la plupart du temps un score *G* inférieur aux modèles précédents et figureront dans le bas du tableau. Ce phénomène peut être atténué en triant le tableau par valeurs croissantes de l'AIC, ainsi les modèles décrivant le plus précisément la distribution de chaque marqueur apparaîtront en premier dans le tableau. En résumé, lors de l'interprétation des modèles multivariés, il faut vérifier si leur AIC est plus bas que celui des modèles univariés correspondant au même marqueur.

La recherche du « meilleur modèle » doit cependant rester parcimonieuse : l'inclusion de nombreuses variables permet d'expliquer précisément la distribution des marqueurs, mais la sur-paramétrisation d'un modèle nuit à l'interprétation écologique des résultats (Schrodt, 2010).

Significativité statistique des modèles logistiques

Le procédure d'estimation des modèles logistiques mise en œuvre par Samβada est assez rigide. Avant de commencer les calculs, l'utilisateur doit choisir s'il veut enregistrer tous les modèles ou seulement ceux qui sont significatifs. Dans le premier cas, il devra filtrer les modèles manuellement après l'analyse (et prend le risque de saturer son disque suivant le nombre de modèles à considérer). Dans le deuxième cas, l'utilisateur indique une p-valeur et Samβada applique les tests de la log-vraisemblance et de Wald avec la correction de Bonferroni : si le seuil de significativité est trop élevé, aucun modèle ne sera détecté. Ces deux cas de figure peuvent se présenter lors de l'analyse de grands jeux de données, ce qui montre les limites de cette approche. Un traitement flexible des modèles logistiques permettrait d'adapter le filtrage des résultats au système étudié. J'entrevois plusieurs options allant dans ce sens.

La première option serait de laisser l'utilisateur choisir entre le test *G*, le test de Wald ou la combinaison des deux. Actuellement, cette option est partiellement intégrée à Samβada : le programme principal applique les deux tests de significativité, alors que le module `Supervision`, chargé du calcul distribué, offre plusieurs options lors de la réunion des résultats. Si l'utilisateur a conservé tous les modèles issus du calcul principal, `Supervision` lui permet de les filtrer selon leurs scores *G* et/ou Wald et lui permet aussi les trier selon leur score *G*, Wald, AIC ou BIC. Ce premier point consisterait donc à homogénéiser les traitements proposés par Samβada et `Supervision`. Cela permettrait d'analyser des jeux de données du type 800k sans avoir à filtrer les modèles manuellement après le calcul.

La deuxième option serait de proposer plusieurs types de corrections du seuil de significativité pour les comparaisons multiples. La correction de Bonferroni contrôle la probabilité de faire au moins une fausse découverte et est bien adaptée aux jeux de données de petite à moyenne taille, lorsqu'un nombre restreint de découvertes est attendu. Or les analyses du génome entier impliquent des millions de tests et de très nombreuses détections sont attendues. Dans ce contexte, le contrôle de la proportion de fausses découvertes permet de détecter des loci au comportement statistiquement significatif alors que ceux-ci auraient été écartés par une

correction du type Bonferroni, tout en limitant le nombre de faux positifs (Storey et Tibshirani, 2003). La FDR selon Storey et Tibshirani est bien adaptée à Samβada si tous les scores *G* et Wald sont enregistrés, notamment car elle ne nécessite pas de se baser sur un modèle significatif obtenu avec la correction de Bonferroni (méthode de Benjamini et Hochberg). Les résultats 800k ont été analysés sur la base du score *G*, mais une FDR selon Storey et Tibshirani permettrait de détecter des modèles selon le test de Wald avec un taux de faux positifs de 5% (table 7.9). Si l'utilisateur choisit la correction FDR et que le nombre de modèles analysés est trop grand pour pouvoir enregistrer tous les scores sur le disque dur pendant le calcul, il serait possible d'en sauvegarder un sous-ensemble. Pour chaque modèle, une procédure aléatoire déterminerait si ses scores doivent être enregistrés, indépendamment de leurs valeurs. La liste des scores enregistrés permettrait d'estimer la distribution des p-valeurs après le calcul. Et les modèles « potentiellement significatifs », c'est-à-dire présentant des scores supérieurs à un seuil estimé *a priori*, seraient également conservés jusqu'à la fin du traitement. Il faut noter que l'algorithme de Storey et Tibshirani calcule les q-valeurs en commençant avec les modèles ayant les plus grandes p-valeurs. Or ces modèles auraient été éliminés durant le calcul pour économiser de la place sur le disque dur. Le calcul des q-valeurs devrait être modifié pour n'utiliser que les modèles enregistrés. Finalement, avant d'appliquer cette méthode à la sélection des modèles multivariés, il faudrait également vérifier qu'elle est compatible avec l'algorithme des « parents » en analysant notamment la distribution des p-valeurs obtenues avec différents jeux de données.

La troisième option concerne les effets des variables environnementales. Si l'utilisateur souhaite comparer les valeurs des paramètres β associés à différentes variables, Samβada doit lui proposer de centrer et réduire ses données environnementales avant de commencer les calculs. Les paramètres β peuvent être corrigés *a posteriori*, mais il est plus simple de modifier les données en amont.

Enfin, la gestion des ressources informatiques pourrait également être améliorée : les ordinateurs actuels comprennent généralement plusieurs coeurs et/ou processeurs. Samβada gagnerait beaucoup de temps en distribuant la charge de calcul entre les coeurs de la machine qu'il utilise [3].

L'implémentation de ces quatre options dans Samβada permettrait d'adapter le traitement des modèles logistiques aux besoins des utilisateurs.

8.2.2 `BayEnv`

Le nombre de SNPs détectés par `BayEnv` ne peut pas être comparé aux résultats des autres méthodes car le seuil de significativité a été fixé de manière arbitraire sur la base des facteurs de Bayes. Néanmoins, `BayEnv` identifie 400 loci parmi les données 54k (soit 1% du total) dont 387 sont également détectés par Samβada. Si les modèles obtenus avec les deux méthodes sont triés en fonction de leur facteur de Bayes ou de leur score *G* (respectivement), l'ordre

3. Des librairies comme `OpenMP` permettent de paralléliser des applications.

des modèles est comparable entre ces méthodes puisque 65 loci sont figurent parmi 100 premiers loci identifiés dans chaque cas. Le type de populations analysées en Ouganda, qui sont composées de nombreux individus hybrides, ne convient peut-être pas à cette méthode, ce qui pourrait expliquer le grand nombre de modèles significatifs identifiés. `BayEnv` est plutôt conçu pour traiter des populations clairement différenciées, du type des 52 populations humaines du HGDD (Coop et al., 2010). De plus, `BayEnv` n'est pas prévu pour analyser si peu de populations différentes (G. Coop, communication personnelle). Il faudrait peut-être n'utiliser que les animaux clairement assignés à une population, comme dans le cas de l'analyse avec `Arlequin`. Dans ce contexte, il est difficile d'interpréter les 2'083 loci (0,8 %) détectés dans les données 800k.

Au delà des résultats obtenus, il faut noter que `BayEnv` est une méthode relativement difficile à utiliser. En effet :

— Elle nécessite de déterminer la structure de population sous-jacente. Cette structure est facile à repérer si les individus sont groupés en populations géographiquement distinctes. A l'inverse, lorsque les individus étudiés sont distribués de manière continue dans l'espace, il faut recourir à une analyse basée sur les loci neutres pour révéler la structure de population.

— Le calcul des corrélations entre populations requiert un sous-ensemble de loci neutres. L'analyse pourrait devenir circulaire au cas où il faudrait fournir *a priori* un ensemble de loci neutres pour identifier les populations et calculer la matrice de corrélation, alors que la méthode est justement conçue pour détecter des signatures de sélection. Cependant, si les populations sont séparées géographiquement, la recherche d'un ensemble de loci neutres peut être évitée en postulant que seule une petite proportion des loci sont soumis à la sélection et en utilisant l'ensemble des données pour calculer la matrice de corrélation.

— Toujours en faisant l'hypothèse que seule une petite proportion des loci est soumise à la sélection, l'estimation de la matrice de corrélation entre populations peut utiliser tous les SNPs. Cette approche évite d'avoir à désigner un ensemble de loci neutres. Toutefois sa faisabilité informatique devrait être testée pour des grands jeux de données.

— Le choix d'une matrice de covariance parmi toutes les estimations induit une incertitude supplémentaire. Coop et al. (2010) propose de faire une moyenne entre plusieurs matrices, mais cette approche n'a pas fonctionné sur ces données. Ce problème est peut-être dû à l'hybridation entre les populations qui est susceptible de produire des instabilités dans l'estimation des corrélations entre populations.

— La création d'un fichier de données pour chaque SNP prend beaucoup de temps et n'est pas efficace informatiquement.

— Le temps de calcul pour détecter les signatures de sélection est également important.

— La recherche des loci sélectionnés parmi les résultats nécessite un traitement particulier. L'approche par la distribution empirique ne permet pas de quantifier le nombre faux

positifs. L'absence d'une estimation de la p-valeur empêche d'appliquer la FDR selon Storey.

— La normalisation des variables environnementales ne permet pas de distinguer leurs effets lorsqu'il y a deux populations. En effet, dans ce cas les deux modèles correspondant à un SNP ont des facteurs de Bayes égaux.

Il convient de préciser que les analyses ont été réalisées avec la première version du logiciel. Günther et Coop (2013) ont depuis développé `BayEnv 2` qui est maintenant disponible sur leur site [4]. Cette version inclut la possibilité de « corriger » les fréquences alléliques pour « compenser » la corrélation de ces fréquences entre populations proches. Cette correction permet d'utiliser des approches basées sur les différenciations entre populations (Beaumont et Nichols, 1996, et dérivés) pour détecter les singularités. `BayEnv 2` permet également de calculer la corrélation de Spearman (ρ) qui moins sensible aux populations « singulières » (par exemple une population très éloignée des autres qui présenterait des fréquences alléliques très différentes à cause de la dérive génétique) que le facteur de Bayes. En effet, lors d'une régression linéaire, quelques événements exceptionnels sont susceptibles de faire apparaître une corrélation globale alors que la plupart des points sont indépendants. La corrélation de Spearman permet de repérer ce phénomène afin d'éviter de tirer des conclusions générales sur des cas particuliers. Cependant, le fonctionnement de `BayEnv 2` tel que décrit par le manuel est similaire à celui de la version utilisée. En particulier, il est toujours nécessaire de créer un fichier par SNP pour détecter les signatures de sélection.

8.2.3 Latent Factor Mixed Model (`LFMM`)

Contrairement à `BayEnv`, le logiciel `LFMM` (Frichot et al., 2013) est assez facile à utiliser. `LFMM` évalue la structure de population en même temps que l'influence de l'environnement. Cette structure n'est cependant pas sauvegardée dans les résultats, ce qui empêche de la comparer avec la structure estimée par une autre approche (par ex. `Admixture`). La principale difficulté d'utilisation consiste à choisir le nombre de populations à considérer. Ce nombre étant généralement estimé à partir d'une étude de la structure de population, il pourrait induire une circularité dans l'analyse. A ce propos, il faut également remarquer que nombre de facteurs latents est généralement plus petit que le nombre de populations estimées par `Admixture` (Frichot, comm. personnelle). Les auteurs suggèrent d'utiliser la méthode de Tracy-Widom pour déterminer le nombre de facteurs latents et mentionnent qu'ils pourraient inclure une validation croisée des résultats (pour déterminer le meilleur K) dans une version ultérieure.

`LFMM` présente l'avantage de pouvoir distribuer les calculs entre les processeurs disponibles, ce qui est bien adapté aux ordinateurs récents qui possèdent plusieurs coeurs. Le traitement prend néanmoins plus de temps qu'avec Samβada.

Au vu des résultats en Ouganda, `LFMM` propose une approche assez conservatrice où 245 loci sont détectés dans les données 54k (soit 0,6%). Il est possible que certains loci soumis

4. gcbias.org/bayenv/

à la sélection aient été inclus dans la structure de population, ce qui expliquerait le peu de SNPs détectés. Le nombre de facteurs latents a été fixé à $K = 1$ en fonction de la structure de population particulière qui oppose zébus et ankoles.

En comparant les résultats de LFMM et de Samβada, si les modèles sont triés en fonction de leur p-valeur, les loci détectés par LFMM ne sortent pas dans le même ordre que ceux détectés par Samβada. Les modèles possédant les plus hauts scores G avec ce dernier ne sont pas détectés par LFMM. Cela pourrait suggérer que les marqueurs les plus fortement corrélés à une variable environnementale sont en fait des SNPs distinguant les deux populations.

Les marqueurs détectés par Samβada et/ou LFMM ont été comparés en fonction de leurs scores et de leur indice d'autocorrélation spatiale (fig. 7.12 p. 130). Ceux détectés conjointement présentent des motifs particuliers de l'AIC et du coefficient à l'origine β'_0 (avec des variables centrées et réduites). Les marqueurs les plus autocorrélés détectés par Samβada ne le sont pas par LFMM. Cette observation est confirmée par la fig. 7.13, où ni les marqueurs les plus autocorrélés, ni ceux ayant les scores G les plus élevés ne sont détectés par LFMM.

Ces informations ne délivrent cependant pas de critère exact permettant de déterminer quelles découvertes de Samβada seront aussi identifiées par LFMM.

8.2.4 Module d'Arlequin pour détecter la sélection naturelle

Le module d'Arlequin permettant de détecter les signatures de sélection naturelle utilise une approche basée sur un modèle hiérarchique en îles (Excoffier et al., 2009 ; Excoffier et Lischer, 2010). Cette approche nécessite d'assigner les individus en populations. Arlequin fournit des résultats très conservateurs : elle détecte 19 loci dans les données 54k, répartis en 13 clusters sur le génome. Seuls trois loci concordent avec les résultats de LFMM, qui se situent dans le même cluster sur le chromosome 5. Treize autres SNPs détectés par Arlequin sont à une distance de moins de 5 millions de paires de base de loci détectés par LFMM.

LFMM et Arlequin fournissent une p-valeur pour chaque modèle, les résultats peuvent donc être filtrés avec la correction de Bonferroni pour les comparaisons multiples. En revanche, la distribution de ces p-valeurs ne permet pas de sélectionner les modèles avec la FDR selon Storey et Tibshirani car l'histogramme de ces p-valeurs ne présente pas la bonne configuration (cf fig. 7.15 p. 135).

8.2.5 Utilité des indices d'autocorrélation spatiale

J'ai analysé la distribution spatiale de trois marqueurs : HM-28 qui est détecté par les quatre méthodes, ARS-94 qui est détecté par Samβada, BayEnv et LFMM, et ARS-11 qui n'est détecté que par Samβada et BayEnv. HM-28 est un allèle plus rare que ARS-11 et ARS-94, situé sur le chromosome 5. On le trouve principalement dans le nord et l'est, avec quelques occurrences à l'extrême sud-ouest.

L'autocorrélation spatiale est souvent décrite comme étant la cause d'associations fallacieuses (Holderegger et al., 2008). Concernant la distribution spatiale de marqueurs génétiques, l'autocorrélation peut être le signe d'une isolation par la distance entre les individus (Meirmans, 2012). Le marqueur autocorrélé présentera alors souvent un gradient de fréquences alléliques qui pourrait être corrélé par hasard avec une variable environnementale. A l'inverse, l'absence d'autocorrélation indique que les individus sont indépendants de leurs voisins, et cette hypothèse est à la base de nombreux tests. Il faudrait vérifier si cette relation entre indépendance spatiale et convergence des approches se confirme avec d'autres marqueurs et dans d'autres conditions, en particulier entre Samβada et `LFMM`. Le cas échéant, la mesure de l'autocorrélation spatiale locale des marqueurs détectés par Samβada permettrait peut-être d'identifier des « candidats privilégiés » dont l'association avec l'environnement n'est sûrement pas fortuite. Comme la création de cartes pour chaque variable nécessiterait trop de temps lorsque de nombreux marqueurs sont détectés, une procédure automatisée pourrait se baser sur le nombre d'association significatives par rapport au nombre de points.

Les marqueurs ARS-11, HM-28 et ARS-94 fournissent quelques informations sur le comportement des différentes méthodes. Le marqueur HM-28 est autocorrélé seulement dans le nord-ouest du territoire et il est détecté par les quatre approches (cf carte 7.19b p. 145). De même, le marqueur ARS-94 est autocorrélé presque uniquement dans le sud-ouest du pays. A l'inverse, ARS-11 est autocorrélé de manière significative sur la moitié du territoire et n'est détecté que par Samβada et `BayEnv`. D'après ces mesures, la différence entre les résultats de Samβada et `LFMM` pourrait s'expliquer par l'intensité de l'autocorrélation spatiale : seuls les loci spatialement indépendants seraient détectés par `LFMM`. Les indices locaux d'associations spatiales bivariées permettent d'étudier la distribution spatiale de la relation entre un marqueur et une variable environnementale. Dans le cas présent, ils ne semblent toutefois pas fournir de critère précis pour distinguer quels marqueurs sont détectés par quelles méthodes. En particulier, le faible nombre de détections d'`Arlequin` n'est probablement pas dû à l'autocorrélation. La carte 7.14 montre que les individus utilisés par « Arlequin » se situent principalement dans le sud-ouest et le centre du pays, c'est peut-être ce sous-échantillonnage qui est à l'origine des résultats obtenus avec cette approche.

Cette étude de l'autocorrélation spatiale est l'occasion de compléter l'analyse du marqueur HM-28. Jusqu'ici nous avons vu qu'il se situe sur le chromosome 5 (p. 138), plus précisément sur un gène qui code pour un facteur de transmission « RFX-4 » (p. 114), qu'il est détecté par les quatre méthodes (avec deux autres loci proches qui sont eux associés à un gène « BT.42818 » impliqué dans la résistance aux bactéries intracellulaires), qu'il est principalement présent au nord-ouest du pays (p. 141) et qu'il est significativement autocorrélé presque uniquement dans cette région (145). En comparant ces deux cartes avec la fig. 8.1, nous pouvons remarquer que la région où ce marqueur est le plus fréquent est également celle où sévit *Trypanosoma brucei gambiense* qui est transmis par la mouche Tsé-tsé et qui provoque la maladie du sommeil chez l'humain. En supposant que la distribution spatiale des deux loci situés près du gène « BT.42818 » soit à peu près semblable à celle de HM-28 (vu qu'ils sont proches), et en supposant également que la région vit *T. b. gambiense* est également touchée par d'autres

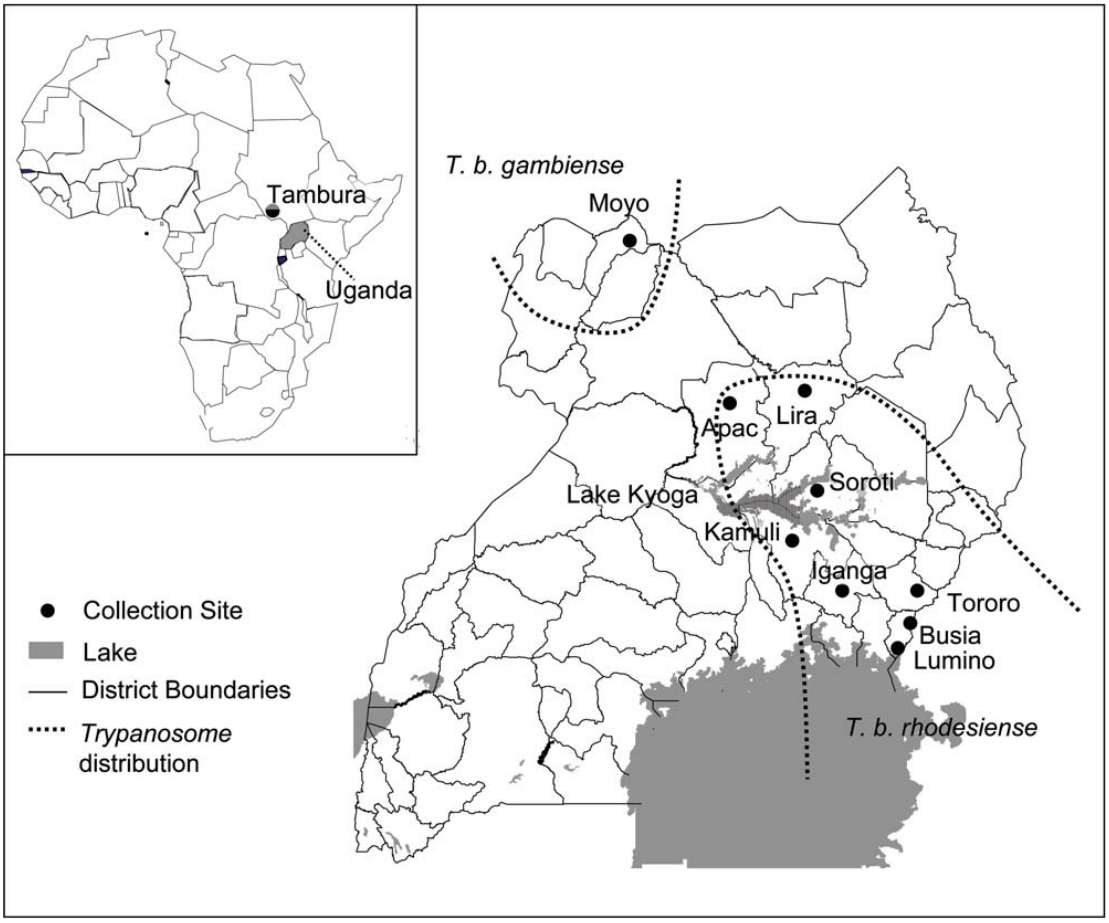


Figure 8.1 – Carte de prévalence de *Trypanosoma brucei gambiense* et *Trypanosoma brucei rhodesiense* en Ouganda. Ces deux vers parasites sont transmis par la mouche Tsé-tsé (*Glossina fuscipes fuscipes*) et provoquent la maladie du sommeil chez l'humain.

trypanosomes affectant les bovins, il est probable que le gène « BT.42818 » soit impliqué dans un processus de résistance aux parasites. Cet exemple montre comment la détection des signatures de sélection naturelle, combinée à une recherche des gènes concernés et à une analyse de la distribution spatiale du marqueur découvert peut nous aider à formuler des hypothèses concernant la fonction biologique associée.

Revenons à l'autocorrélation spatiale des SNPs détectés. Dans l'idéal, Samβada devrait être capable de fournir des statistiques spatiales pour tous les modèles, mais cela prendrait trop de temps pour les grands jeux de données. Cependant, cette étude suggère que l'utilisation de l'indice LISA de concert avec un score de significativité permet d'écarter les loci trop autocorrélés et qui sont probablement plutôt des signatures démographiques.

L'analyse spatiale fournit des approches pour étudier les processus où les observations dépendent de leurs voisines géographiques. Comme nous l'avons vu précédemment, l'autocorrélation spatiale locale pourrait permettre de distinguer parmi les loci détectés par Samβada ceux qui sont aussi identifiés par `LFMM`. Par contre les détections d'`Arlequin` semblent plus difficiles à expliquer (cf sec. 7.1.8 et ci-dessus). La régression géographiquement pondérée (*Geographically Weighted Regression, GWR*) est une méthode d'analyse spatiale spécialement prévue pour modéliser des relations qui varient dans l'espace (Fotheringham et al., 2002). La GWR produit une régression par point, dont les coefficient sont calculés à partir des points voisins. Les coefficients sont des fonctions continues des coordonnées géographiques, mais ces fonctions n'ont pas besoin d'être déterminées avant l'analyse. La relation entre les variables explicatives et la variable prédite peut être linéaire, logistique ou de Poisson et la fonction de pondération est définie comme dans le cas de l'autocorrélation spatiale. L'analyse de modèles GWR multivariés permet de déterminer quels sont les coefficients qui varient dans l'espace et ceux dont la valeur est fixe. La GWR permet de tenir compte de l'autocorrélation qui produit des faux positifs dans les modèles non-spatiaux.

8.2.6 Approches comparées

La comparaison des résultats fournis par Samβada, `BayEnv`, `LFMM` et `Arlequin` montre que Samβada est une méthode rapide mais qu'elle détecte beaucoup de marqueurs dont certains sont certainement des faux positifs. A l'inverse, `LFMM` et `Arlequin` nécessitent plus de temps de calcul mais utilisent des approches conservatrices et détectent moins de marqueurs. `BayEnv` est la méthode la plus lente et l'analyse de ses résultats est rendue difficile par le choix empirique du seuil de significativité.

Les méthodes différent également par leurs pré-requis. Samβada n'en a aucun et utilise des données individuelles, ce qui constitue un avantage ; `BayEnv` nécessite un jeu de loci neutres ainsi qu'une répartition des individus en populations ; `Arlequin` n'utilise que des individus clairement assignés à des populations ; `LFMM` doit connaître le nombre de populations présentes dans l'échantillon.

La principale caractéristique des bovins ougandais est certainement leur structure de population due à l'hybridation des espèces ankole et zébu. Les ankoles n'ont cependant qu'une petite partie de leur génome provenant des zébus alors que ceux-ci ont souvent une part importante de leur patrimoine génétique d'origine ankole. La répartition des populations est corrélée à la latitude, ce qui crée des associations fortuites entre la fréquence de certains marqueurs génétiques et l'environnement (facteur de confusion). Ceci explique pourquoi Samβada, qui ne tient pas compte de la structure de population, détecte beaucoup plus de loci potentiellement soumis à la sélection que `LFMM` et `Arlequin`. Ces dernières sont des méthodes conservatrices, dont les résultats ne concordent que pour les trois marqueurs communs à toutes les approches. Ces marqueurs sont situés dans la même région du chromosome 5. Samβada les identifie aussi dans plusieurs analyses au sien desquelles très peu d'autres marqueurs sont détectés (dans les données 800ksub, dans la population zébu et dans les modèles bivariés incluant la variable « ankole »). La comparaison entre la distribution spatiale d'un de ces marqueurs et celle d'une mouche Tsé-tsé transmettant la maladie du sommeil à humain suggère que le gène « BT.42818 », proche de ces marqueurs sur le chromosome, pourrait être impliqué dans un processus de résistance aux parasites.

La structure de population n'est pas la seule source de fausses découvertes lors de la détection de signatures de sélection naturelle. Les différents types de dynamique démographique, comme la croissance ou la contraction, l'effet fondateur lors d'une expansion territoriale ou les goulots d'étranglement (réduction soudaine de la population, *bottleneck*) produisent également des facteurs de confusion (Li et al., 2012). Dans le contexte des grands jeux de données actuels, les marqueurs génétiques sont si nombreux qu'ils se retrouvent très proches sur le génome et sont corrélés à leurs voisins. En effet, ils sont souvent transmis ensemble lors du processus de reproduction et ne peuvent plus être considérés comme indépendants. L'analyse statistique de ce déséquilibre de liaison est utilisée dans certaines approches très récentes en génomique des populations (Jensen et al., 2007). Ces méthodes permettent de tenir compte de l'histoire démographique d'une population mais requièrent de nombreux calculs. Elles n'ont pas été intégrées à ce présent travail qui vise à développer une approche rapide en génomique environnementale.

Cette étude montre que Samβada est une approche rapide sans pré-requis qui détecte beaucoup de loci alors que `LFMM` et `Arlequin` nécessitent des informations préalables sur les organismes étudiés mais sont bien plus conservatrices. Il est cependant difficile de déterminer exactement quels loci fournissent un avantage adaptatif sur la base de ces analyses. La génomique environnementale et les approches basées sur les singularités fournissent des informations sur les loci potentiellement soumis à la sélection. Ces candidats doivent ensuite être validés par des expériences mesurant leur effets physiologiques ou phénotypiques. Une étude fonctionnelle de ces loci permettrait de déterminer leur rôle biologique et d'identifier leur avantage sélectif (Holderegger et al., 2008). Toutefois, les bases de données génomiques en ligne permettent de retrouver les marqueurs identifiés et de déterminer s'ils se situent près d'un gène dont la fonction a parfois déjà été identifiée. Il peut également s'agir d'un gène « potentiel » dont la présence a été prédite sur la base d'une séquence d'ADN typique de

l'extrémité d'un gène ou alors qui est similaire à un gène déjà identifié chez une autre espèce. La découverte d'une signature de la sélection naturelle dans cette région pourrait permettre d'y confirmer la présence d'un gène.

8.2.7 L'apport des données simulées

Les données simulées avec CDPOP constituent une tentative d'imiter les données empiriques afin de distinguer le comportement des loci soumis à la sélection de celui des loci neutres. Les jeux de données empiriques et simulés sont difficilement comparables. Ils diffèrent d'abord par leur taille : la population simulée avec CDPOP comporte 100 SNPs. Cette configuration n'est pas susceptible de reproduire le manque de puissance de Samβada avec les données 800k. De plus, les individus simulés peuvent parcourir jusqu'à un quart de la distance maximale pour trouver un partenaire, alors que la plus grande distance du sud-ouest au nord-est de l'Ouganda est d'environ 1'000 km : hormis lors du cas d'un achat d'animal à l'étranger, il est peu probable qu'un éleveur parcoure 250 km pour acquérir un taureau. Il faut cependant remarquer qu'il pourrait y avoir des déplacements de vaches à l'intérieur du pays sous l'impulsion des associations d'éleveurs.

Les portées sont également de trop grande taille pour des bovins ($\lambda = 4$), et les coefficients de sélection ne reflètent pas vraiment la situation réelle. La sélection faible (1% de mortalité à la naissance) est trop basse pour créer une différentiation entre le nord et le sud du terrain, et l'ampleur des déplacements autorisés gomme toute trace de la sélection. La sélection forte (50% de mortalité à la naissance) est trop élevée, car les individus subissant cette pression en Ouganda n'auraient pas produit suffisamment de descendants pour que leur lignée survive. La sélection modérée (10% de mortalité à la naissance) est plus vraisemblable, mais les déplacements autorisés annulent une partie de cette pression de sélection. Les vaches ougandaises présentent un gradient de population nord-sud qui est en partie dû à l'isolation par la distance entre les différentes régions.

Bien qu'Admixture détecte deux populations dans les données simulées, elles sont issues de la même population ancestrale qui n'a jamais été scindée en deux groupes. Les zébus et les ankoles ont longtemps évolué séparément avant de se retrouver en Ouganda. Leur interaction est celle de deux populations distinctes qui se mélangent, pas d'une population qui se différencie sous l'effet de la sélection. Une simulation des bovins ougandais avec CDPOP devrait tenir compte de cette configuration.

La comparaison des résultats de Samβada et LFMM montre qu'ils détectent les même marqueurs quand la sélection est moyenne ou forte. Pour la sélection faible, Samβada trouve trois faux positifs alors que LFMM ne les détecte pas. Ces résultats suggèrent que Samβada et LFMM détectent les même signatures si la sélection est forte, alors que Samβada détecte quelques faux positifs quand la sélection est faible.

L'analyse de ces données fournit également des informations sur l'étude de marqueurs domi-

nants avec Samβada. Tous les loci ont été traités comme des SNPs, car le caractère dominant de certains marqueurs ne peut pas être déterminé a priori. Lorsque le locus adaptatif est identifié par Samβada, c'est son allèle GG (et parfois AA) qui est détecté. Or les allèles GG sont l'inverse du génotype dominant AA+AG. La méthode est donc capable d'identifier les marqueurs dominants soumis à la sélection dans le cas de SNPs biallélliques. Cela signifie qu'une signature de sélection pour un génotype homozygote pourrait être le signe que le génotype inverse est dominant et sélectionné. En outre, la comparaison des histogrammes de l'autocorrélation spatiale globale entre les différents scénarios de sélection montre que les allèles GG voient leur autocorrélation augmenter avec la pression de sélection. *Cet effet permet peut-être d'identifier les loci dominants soumis à la sélection dans les données empiriques.*

En résumé, les simulations numériques offrent la possibilité de créer des populations aux caractéristiques connues. Elles peuvent servir à tester les méthodes de détections de loci soumis à la sélection naturelle dans un environnement contrôlé, à condition que le système ait été calibré pour reproduire au mieux l'habitat et le comportement des organismes étudiés.

La section suivante présente quelques pistes pour éviter les fausses découvertes liées à la structure de population avec Samβada.

8.3 Intégration de la structure de population dans Samβada

Plusieurs approches peuvent être envisagées pour limiter le nombre de fausses découvertes liées à la structure de population dans Samβada.

La première serait d'utiliser ce logiciel pour pré-identifier des marqueurs potentiellement sous sélection naturelle puis de les analyser avec une autre approche plus conservatrice : les méthodes détectant les singularités, comme `Arlequin`, sont basées sur l'hypothèse que les loci adaptatifs sont beaucoup moins nombreux et plus (ou moins) différenciés (F_{ST}) que les loci neutres. Il faudrait élaborer une méthode de ce type acceptant un ensemble de loci connus pour être neutres et un ensemble de loci à tester. `BayEnv` est une approche corrélative basée sur ce principe, mais elle semble plutôt adaptée à des études avec beaucoup de populations clairement séparées (Coop et al., 2010). Deux méthodes pourraient convenir, mais il faudrait tester leur comportement dans un contexte où les loci neutres sont moins nombreux que les candidats à la sélection : `LFMM` en génomique environnementale et `BayeScan` (Foll et Gaggiotti, 2008) en génétique des populations. Il n'est pas certain que cette dernière soit utilisable, car son application aux bovins ougandais n'a pas fonctionné (Orozco-terWengel, communication personnelle).

La deuxième possibilité serait d'intégrer la structure de population dans dans une analyse multivariée avec Samβada. Cette structure pourrait être modélisée par des *Moran's eigenvector maps* qui permettent de tenir compte des variables inconnues (cf p. 22 et Manel et al., 2010). Les résultats d'une analyse de populations, par ex. avec `Admixture`, permettent également de tenir compte de cette structure. Cette méthode s'est révélée très conservatrice et a identifié

des loci également détectés avec `LFMM` et `Arlequin`. Il faudrait tester si elle fonctionne dans d'autres contextes et si un test de significativité basé sur la FDR selon Storey et Tibshirani pourrait identifier plus de loci concordant avec d'autres approches.

L'analyse spatiale fournit des approches pour étudier les processus où les observations dépendent de leurs voisines géographiques. Dans le cadre des approches corrélatives, la mesure de l'autocorrélation spatiale locale aide à interpréter les résultats. L'idéal serait d'arriver à modéliser l'adaptation locale en même temps que l'autocorrélation spatiale La régression géographiquement pondérée (*Geographically Weighted Regression, GWR*) pourrait fournir le moyen de distinguer les loci sous sélection naturelle de ceux qui sont impliqués dans la structure de population.

Ces différentes approches doivent être testées avec des modèles logistiques dans plusieurs études de génomique environnementale, en variant notamment à l'étendue géographique de la région concernée. La méthode permettant de séparer les effets de la sélection de ceux de la structure de population sera intégrée à Samβada.

8.4 Diffusion de Samβada

La génomique environnementale prend racine en écologie du paysage, en analyse spatiale et en génétique des populations. Elle requiert maintenant également clairement des notions avancées en informatique. C'est une approche multidisciplinaire qui doit être facile à mettre en œuvre car elle met en relation des chercheurs de différents horizons. Les principaux obstacles à sa diffusion sont l'accès aux données environnementales, la puissance de calcul nécessaire et l'interprétation des résultats.

L'accès aux variables environnementales peut être facilité de deux manières. De nombreuses bases de données topo-climatiques sont désormais accessibles en ligne. Ces données peuvent être extraites et préparées avec des logiciels SIG libres. La mention explicite de ces outils lors de la présentation des méthodes dans les publications et les conférences les feraient connaître des chercheurs sans spécialisation en analyse spatiale. La préparation de marches à suivre pour l'extraction des variables environnementales nécessaires aux approches corrélatives aiderait à ces chercheurs à démarrer leurs analyses.

Une autre solution est de mettre à leur disposition une plate-forme en ligne (sur Internet) pour la génomique environnementale. Cette approche est en cours de développement : `GEOME` est un site Internet combiné à une base de donnée géographique (Emery, 2012), sur laquelle les utilisateurs peuvent charger leurs données génétiques géo-référencés (la taille des fichiers est limitée). `GEOME` extrait les données environnementales qu'il possède pour ces individus et appelle Samβada en arrière-plan. Dès que le traitement est terminé, l'utilisateur reçoit un courriel et peut récupérer ses résultats. La principale limitation de `GEOME` est qu'il fonctionne pour l'instant sur une seule machine et doit répartir sa capacité de calcul entre les utilisateurs. Ceux-ci ont néanmoins la possibilité de télécharger leur fichier de données environnementales

pour les analyser eux-mêmes.

Le deuxième obstacle à la démocratisation de la génomique environnementale est la disponibilité de moyens informatiques appropriés. De ce point de vue, Samβada a l'avantage d'être facile à paralléliser : les calculs concernant chaque marqueur génétique sont indépendants et peuvent être traités séparément. Les calculs peuvent ainsi être répartis sur plusieurs machines, pas forcément identiques et pas forcément synchronisées par un système de gestion de tâches.

Concernant les résultats et leur interprétation, Samβada cherche à fournir des informations aussi claires que possible sur la façon dont il fonctionne et sur les éventuelles difficultés qui peuvent se présenter. Les prochaines étapes concernent la possibilité de trier les résultats selon plusieurs critères et de les filtrer selon plusieurs seuils de significativité après le calcul. La simplification de l'utilisation passe également par l'homogénéisation des fichiers de paramètres utilisés par Samβada et `Supervision`. Le format des résultats de l'autocorrélation spatiale doit aussi être revu : Samβada fournit des fichiers compatibles avec les principaux programmes SIG, mais ces données doivent être mises en forme manuellement. L'interprétation serait facilitée si les résultats étaient fournis avec des fichiers de paramètres destinés par exemple à `QuantumGIS` [5] qui permettraient d'afficher directement les résultats selon un thème standard. L'utilisation de Samβada serait également facilitée s'il était distribué avec un plug-in permettant de l'appeler directement depuis l'interface graphique de `QuantumGIS`. Cela permettrait aux utilisateurs d'extraire leurs variables environnementales, de lancer leurs calculs et d'analyser l'autocorrélation spatiale avec le même logiciel.

La résolution de ces difficultés permettra de favoriser l'utilisation de cette méthode par le plus grand nombre, notamment dans le cadre de programmes de conservation de la biodiversité.

5. logiciel SIG libre

9 Conclusion

L'avènement du séquençage intégral du génome a révolutionné notre approche de la génétique. NextGen est le premier projet en génétique de la conservation à étudier l'adaptation locale à l'environnement en séquençant 320 animaux de deux espèces différentes à l'échelle d'un pays sur la base d'un tel jeu de données [1]. Ce volume d'information inédit requiert l'application de nouvelles méthodes d'analyse. NextGen est ainsi à l'origine de Samβada, logiciel de génomique environnementale réunissant détection de signatures de sélection naturelle et analyse spatiale des régions du génome en question. Ce projet a également fourni à Samβada son premier grand jeu de données avec l'étude des bovins ougandais. Cet épilogue s'ouvre sur une synthèse des caractéristiques de Samβada.

9.1 Contribution globale de Samβada en génomique environnementale

L'analyse des résultats obtenus chez les vaches ankoles et les zébus en Ouganda nous fournit des éléments de réponses aux questions soulevées au début de ce travail (voir p. 13).

Modèles univariés

L'étude de l'adaptation locale chez les bovins ougandais à partir de deux jeux de données génotypiques à moyenne et haute densité (50k et 800k SNPs) a démontré la capacité de Samβada à détecter les signatures de la sélection naturelle dans un contexte génomique. En effet, Samβada est une méthode rapide, apte à traiter de grands jeux de données, et qui ne requiert comme données d'*input* qu'une description environnementale des habitats des organismes génotypés. Dans le contexte logistique univarié, le traitement de chaque marqueur étant indépendant des autres, la charge de calcul est facile à répartir entre plusieurs ordinateurs. Samβada est ainsi polyvalent et est utilisable aussi bien sur un ordinateur personnel que sur une grappe de calcul. Sa simplicité le rend toutefois susceptible de faire de fausses détections si

1. nextgen.epfl.ch

les organismes étudiés présentent une structure de population. (L'effet démographique mime l'effet de la sélection, Holderegger et al., 2008). **Lorsque les approches corrélatives prennent correctement en compte cette structure, les résultats obtenus se rapprochent de ceux produits par les méthodes de génomiques des populations** [2] (cf question 4 p. 13). Dans ce cas, la génomique environnementale présente l'avantage de permettre formuler des hypothèses écologiques sur les processus adaptatifs observés.

Modèles multivariés

L'analyse des modèles multivariés permet, d'une part, d'étudier l'effet d'une combinaison de variables environnementales et de déterminer ainsi si la conjonction de plusieurs prédicteurs permet une meilleure description de la distribution d'un marqueur. Il convient cependant de favoriser la parcimonie des modèles multivariés, en particulier de vérifier si l'information supplémentaire obtenue justifie d'augmenter la complexité du modèle. D'autre part, **l'analyse multivariée permet d'intégrer au besoin une variable représentant la structure de population** (cf. question 2 p. 13) si cette information est pertinente pour l'organisme étudié. Nous pouvons remarquer à ce propos que la structure de population est susceptible de constituer un facteur de confusion, par exemple lorsque que la région d'étude est beaucoup plus grande que la distance de dispersion typique des individus. En revanche les études menées localement sont moins susceptibles d'être confrontées à ce problème.

Statistiques spatiales

Samβada peut mesurer l'autocorrélation spatiale locale et globale des marqueurs génétiques, ce qui permet de visualiser les régions où les fréquences alléliques se ressemblent ou diffèrent. Dans les régions où l'autocorrélation spatiale n'est pas significative, la présence d'un marqueur chez un individu est indépendante du génotype de ses voisins. Ces zones sont particulièrement importantes, car la régression logistique utilisée par Samβada ne tient pas compte des ressemblances entre voisins, qui peuvent mener à de fausses découvertes. La détection d'une signature de sélection pour un marqueur ne présentant pas ou peu d'autocorrélation spatiale démontre que la relation sous-jacente n'est pas due à une corrélation fallacieuse [3]. D'un point de vue pratique, l'utilisation de l'autocorrélation locale pour valider les détections nécessiterait d'améliorer la vitesse du calcul et de trouver un critère pour identifier automatiquement les marqueurs dignes d'intérêt ; la cartographie manuelle de tous les marqueurs détectés est en effet trop fastidieuse pour être efficace. L'exploitation des statistiques spatiales repose donc en l'état actuel sur l'application d'une démarche en deux temps, soit 1) l'identification des marqueurs les plus significatifs, et 2) le calcul des statistiques spatiales sur ce sous-groupe restreint.

2. Fdist dans `Arlequin` par exemple.
3. Cf marqueur HM-28, détection commune aux quatre méthodes, p. 148.

Stratégie d'échantillonnage

L'étude de l'adaptation locale en génomique environnementale est très sensible à la stratégie d'échantillonnage adoptée. Les habitats étudiés doivent premièrement être aussi distincts que possibles pour observer la présence des marqueurs dans diverses conditions. De nombreuses méthodes, comme Samβada et `LFMM`, utilisent l'individu comme unité d'analyse, ce qui rend possible une modélisation précise de la présence des marqueurs en fonction de l'environnement, mais ce qui les rend aussi sensibles à la distribution spatiale des points d'échantillonnage (voir p. ex. les facteurs de confusion dus à l'autocorrélation spatiale et aux pseudo-réplicats). Dans ce contexte, l'analyse statistique multivariée permet de caractériser les différents habitats présents dans la zone d'étude. Une analyse préalable permet d'organiser un échantillonnage stratifié qui fournira une information optimale sur la zone étudiée. Lorsque cette approche n'est pas applicable, ou lorsqu'il faut déterminer quels échantillons seront séquencés, **ce type d'analyse permet de sélectionner un sous-ensemble d'échantillons représentant tous les habitats identifiés tout en maximisant la distribution spatiale des individus concernés** (cf question 5 p. 13). Les résultats des futures analyses sur les petits ruminants au Maroc nous en feront certainement la démonstration !

Favoriser l'utilisabilité

La diffusion des approches corrélatives auprès des chercheurs sans spécialisation en bioinformatique est confrontée à trois défis : l'accès aux données environnementales, la facilité d'utilisation des méthodes (y compris le formatage des données) et l'interprétation des résultats. Samβada se concentre sur les deux derniers points, en proposant **des modules informatiques qui permettent de traduire les données moléculaires à partir de formats courants, et en essayant de fournir une gamme d'options utiles et faciles à paramétrer tout en fournissant des résultats filtrés et triés ainsi que des valeurs d'autocorrélation locale directement importables dans des logiciels SIG** (cf question 6 p. 13). L'étape suivante consiste à fournir des méthodes de cartographie automatique des résultats, voire de simplification de l'extraction de variables environnementales et d'utilisation de Samβada avec une interface graphique.

9.2 Apports respectifs des approches utilisées

Les résultats obtenus en Ouganda suggèrent que les études menées à grande échelle géographique exigent la prise en compte de la structure de population le cas échéant, étant donné que l'analyse univariée menée avec Samβada détecte beaucoup plus de signatures de sélection que `LFMM` ou `Arlequin`, qui intègrent cette structure. Cette observation est cohérente avec des études précédentes (De Mita et al., 2013).

Toutefois, dans un contexte de séquençage du génome entier, le nombre de marqueurs détectés est appelé à augmenter avec la taille des jeux de données considérés. La proportion des loci détectés par Samβada ($\sim$ 6%) est en effet comparable à des études menées sur de

Méthode	**Famille de méthode**	**Exige des populations clairement identifiées**	**Prérequis**	**Vitesse de traitement**
Samβada GLM	Corrélative	Non	Aucun	Rapide +
Samβada GLM avec SP	Corrélative	Non	Coefficients d'appartenance aux pops	Rapide
Samβada AS	Stats. spatiales	Non	Position des individus	Lente
`BayEnv`	Corrélative	Oui	Ensemble loci neutres + Ind. groupés par pop	Lente
`LFMM`	Corrélative	Non	Nb de facteurs latents (K)	Rapide
`Arlequin`	GPop	Oui	Ind. groupés par pop.	Rapide

Méthode	**Utilisation**	**Comportement de la détection**	**Evaluation par…**	**Remarques**
Samβada GLM	Facile +	Libéral	p-valeurs	Sensible à la SP
Samβada GLM avec SP	Facile +	Conservateur	p-valeurs	
Samβada AS	Facile	-	p-valeurs	Implique schéma de pondération, traitement manuel des résultats
`BayEnv`	Difficile	Conservateur (si pops. distinctes)	Facteurs de Bayes	Conçu pour beaucoup de pops clairement séparées
`LFMM`	Facile	Conservateur	p-valeurs	Estime la SP en même temps que l'effet des var. env., nécessite de choisir K
`Arlequin`	Facile	Conservateur	p-valeurs	Doit couper les grands jeux de données en morceaux

Table 9.1 – Vue d'ensemble des méthodes utilisées. Abbréviations : GPop = génétique des populations, SP = structure de population, AS = mesure de l'autocorrélation spatiale.

plus petits jeux de données (Poncet et al., 2010 ; Manel et al., 2010). De plus les résultats de `LFMM` et d'`Arlequin` sont très conservateurs ($\sim 0,1\%$ de loci détectés) et il est possible que ces méthodes « oublient » des faux négatifs dans leur analyse. Bien que `BayEnv` tienne compte de la structure de population, cette méthode ne permet pas d'utiliser les tests de significativité usuels, ce qui complique la comparaison de ses résultats avec ceux des autres méthodes. Même en tenant compte du fait que `BayEnv` propse une technique différente pour sélectionner les modèles, ces résultats suggèrent que cette méthode n'est pas adaptée aux populations qui se mélangent (cf sec. 8.2.2).

Une approche publiée récemment, `SGLMM` (Guillot et al., 2013) n'a pas pu être incluse dans les analyses des bovins ougandais. Il sera intéressant de l'utiliser dans des études ultérieures en génomique environnementale afin de comparer ses résultats avec les autres méthodes dont la table 9.1 présente un résumé des caractéristiques principales et la table 9.2 fournit quant à elle des indications utiles sur les méthodes utilisées dans cette thèse en fonction du contexte d'analyse.

Méthode	**Taille des données moléculaires (loci)**	**Populations**	**Données**	**Très grand volume de données**
`BayEnv`	Petite ($\rightarrow$ 100k)	Distinctes	Par pop.	Très long, gestion des fichiers problématique
`Arlequin`	Moyenne ($\rightarrow$ 1M)			Peut traiter les données moléculaire par blocs
`LFMM`		Mélangées	Par indiv.	Peut traiter chaque variable env. séparément, calcul multitâche
Samβada	Grande ($\rightarrow$ 1M, bivarié)			Peut traiter les données moléculaires par blocs, module de parallélisation inclus
`CoreSAM`	Grande ($\rightarrow$ 20M, univarié)			

Table 9.2 – Vue synoptique des approches applicables en fonction du type de traitement. Des données individuelles peuvent être agrégées par populations, l'opération inverse est rarement possible. Les recommandations sur les tailles des jeux de données sont des estimations basées sur les analyses présentées dans ce travail.

9.3 Séquençage intégral et stratégie d'analyse

Le volume de données produit par le séquençage intégral nécessite un traitement efficace. Samβada est capable d'**analyser rapidement de grands jeux de données** (cf objectif 1 p. 13), mais est sujet à la détection de faux positifs. Or si le taux de détections reste aux alentours

de 5%, un jeu de 20 millions de marqueurs mènerait potentiellement à environ 1 million de loci détectés, ce qui demanderait de toute manière de trier les modèles les plus intéressants. En effet, les loci détectés sont des candidats dont le rôle biologique doit être vérifié par des tests fonctionnels, qui prennent beaucoup de temps[4]. `Arlequin` et `LFMM` pourraient éventuellement être appliqués dans ce contexte, en utilisant une infrastructure de calcul très puissante. De par sa gestion des fichiers de données, `BayEnv` n'est clairement pas applicable dans ce cas.

Une première solution consisterait à utiliser Samβada pour sélectionner un sous-ensemble de marqueurs potentiellement soumis à la sélection, puis de soumettre ces marqueurs à une autre méthode d'analyse, plus robuste mais plus lente. Le principal écueil est que les méthodes de détection de sélection naturelle partent généralement du principe que la sélection est un évènement rare et que la plupart des loci ont un comportement neutre. Avant de recommander cette méthode, il faudrait vérifier la fiabilité des résultats fournis par `LFMM` et `Arlequin` si le jeu de données qu'ils traitent contient beaucoup de loci sous sélection.

Les fonctionnalités d'analyse spatiale de Samβada suggèrent une autre stratégie. En effet, **la distribution spatiale de l'autocorrélation locale est capable d'indiquer quels modèles significatifs sont les plus intéressants** (cf question 3 p. 13). Cette méthode est *a priori* applicable, mais le critère de choix reste à identifier, ce qui nécessite de plus amples investigations. La première étape serait de sélectionner les marqueurs présentant un minimum d'autocorrélation locale et de vérifier s'ils sont détectés par d'autres méthodes. Une autre façon d'appliquer cette stratégie consisterait à tester si des régressions géographiquement pondérées (GWR), méthode qui intègre l'autocorrélation spatiale, appliquées au jeu initial de marqueurs détectés permettraient de filtrer les faux positifs.

La troisième stratégie proposée est d'estimer la structure de population avec une méthode adaptée aux grands jeux de données[5] et d'utiliser les coefficients d'appartenance à certaines populations comme variables explicatives dans Samβada. Cette approche a fourni ici des résultats très conservateurs en ne détectant que les loci communs à `LFMM` et `Arlequin` (cf sec. 7.1.3). Avant d'appliquer cette approche aux données issues du séquençage du génome entier, il faudrait tester d'une part son comportement en utilisant une sous-sélection des modèles basée sur le FDR selon Storey et Tibshirani plutôt que sur la correction de Bonferroni. En effet, il faudrait vérifier si ce FDR est compatible avec la sélection des modèles basée sur les parents et aussi si les modèles détectés par ce moyen sont de vraies découvertes. D'autre part, le calcul des modèles bivariés prenant beaucoup de temps, il faudrait laisser le choix à l'utilisateur de ne calculer que certains des modèles bivariés (par exemple ceux incluant la ou les variable(s) décrivant la structure de populations). Cette approche peut être très rapidement mise en œuvre.

4. Les loci identifiés par des méthode de détection de singularités (comme `Arlequin`) sont aussi concernés.
5. `sNMF`, *sparse Non-negative Matrix Factorisation*, est par exemple prévue pour calculer rapidement les coefficients d'appartenances aux populations (http://membres-timc.imag.fr/Eric.Frichot/snmf/index.htm).

9.4 Perspectives

D'un point de vue théorique, la recherche en génomique environnementale devrait se concentrer sur le développement d'une approche commune avec la génomique des populations basée sur des modèles théoriques de génétique des populations (Joost et al., 2013). En effet, la mise au point d'une méthode permettant de détecter rapidement les signatures de sélection naturelle tout en identifiant les facteurs de confusion que sont la structure de population et l'histoire démographique constituerait une avancée majeure dans notre compréhension des processus génétiques sous-tendant l'adaptation locale.

Dans cette étude, les détections les plus significatives correspondent entre les différentes méthodes. Par contre les autres marqueurs identifiés diffèrent d'une méthode à l'autre. Ceci tend à montrer que l'utilisation parallèle de plusieurs méthodes est malheureusement toujours la manière la plus robuste pour identifier les régions du génome sous sélection. Des expérimentations plus poussées permettraient de mieux comprendre les comportements respectifs des méthodes et leurs relations mutuelles. Etant donné que le déséquilibre de liaison peut également servir à détecter la sélection dans des populations de taille variable (Jensen et al., 2007), il serait intéressant d'intégrer ce type d'approche dans des travaux ultérieurs.

D'un point de vue pratique, il est nécessaire de trouver des vecteurs pour diffuser les résultats des études auprès d'un public non-spécialiste. Par exemple, les gestionnaires en charge de la conservation de la biodiversité ont besoin de synthèses des études réalisées dans leur pays (Bruford, 2011). Des plate-formes comme `GEOME` (Emery, 2012) ou comme *ConGRESS*[6] à plus large échelle peuvent être utiles à des experts en conservation sans spécialisation en (bio)informatique, ou à jeter des ponts entre les chercheurs et le grand public.

6. http ://www.congressgenetics.eu/

A Base de données des échantillons

Le suivi de la campagne d'échantillonnage a été facilité par la mise en place d'une base de données accessible via une interface Web (voir sec. 4.1.3 p. 30). La création de cette plateforme s'est appuyé sur `EasyDev` (Mingard, 2008). Cet environnement de développement permet, d'une part, de générer une base de données `MySQL` pour sauvegarder l'information et, d'autre part, de créer les objets informatiques nécessaires au serveur Web pour interagir avec cette base de données dans le langage PHP [1]. `EasyDev` assure la gestion technique du projet, ce qui permet de concentrer le développement sur la structure logique et sur l'interface avec l'utilisateur.

A.1 Structure de la base de données

L'échantillonnage étant centré sur les individus, la base de données reprend ce principe et s'organise autour de la table répertoriant les animaux (voir figure A.1). Chaque animal possède un numéro, un nom et une position géographique (ainsi que d'autres attributs) et est relié à une espèce, à une race, à un pays et à une cellule d'échantillonnage (voir figure A.1 note 1). La table des utilisateurs sous-tend également la structure de la base de données (voir note 2). Chacun d'entre eux appartient à une organisation et est rattaché à un groupe d'utilisateurs (**user_type**), ce qui détermine ses droits d'édition dans la base de données (relation **perm** avec la table **action**). Le suivi de la campagne est réalisé avec la table **sampling_progress** (voir note 3) qui répertorie les cellules où l'échantillonnage d'une espèce est terminé.

A.2 Interface Web

Les partenaires NextGen accèdent à la base de données via un site Web construit autour des classes PHP fournies par `EasyDev`. La figure A.2 présente trois utilisations du site pour la saisie des échantillons (a), le suivi de la campagne (b) et la validation des cellules terminées (c).

1. Langage de programmation dédié à la création de pages Web dynamiques via un serveur HTTP (www.php.net).

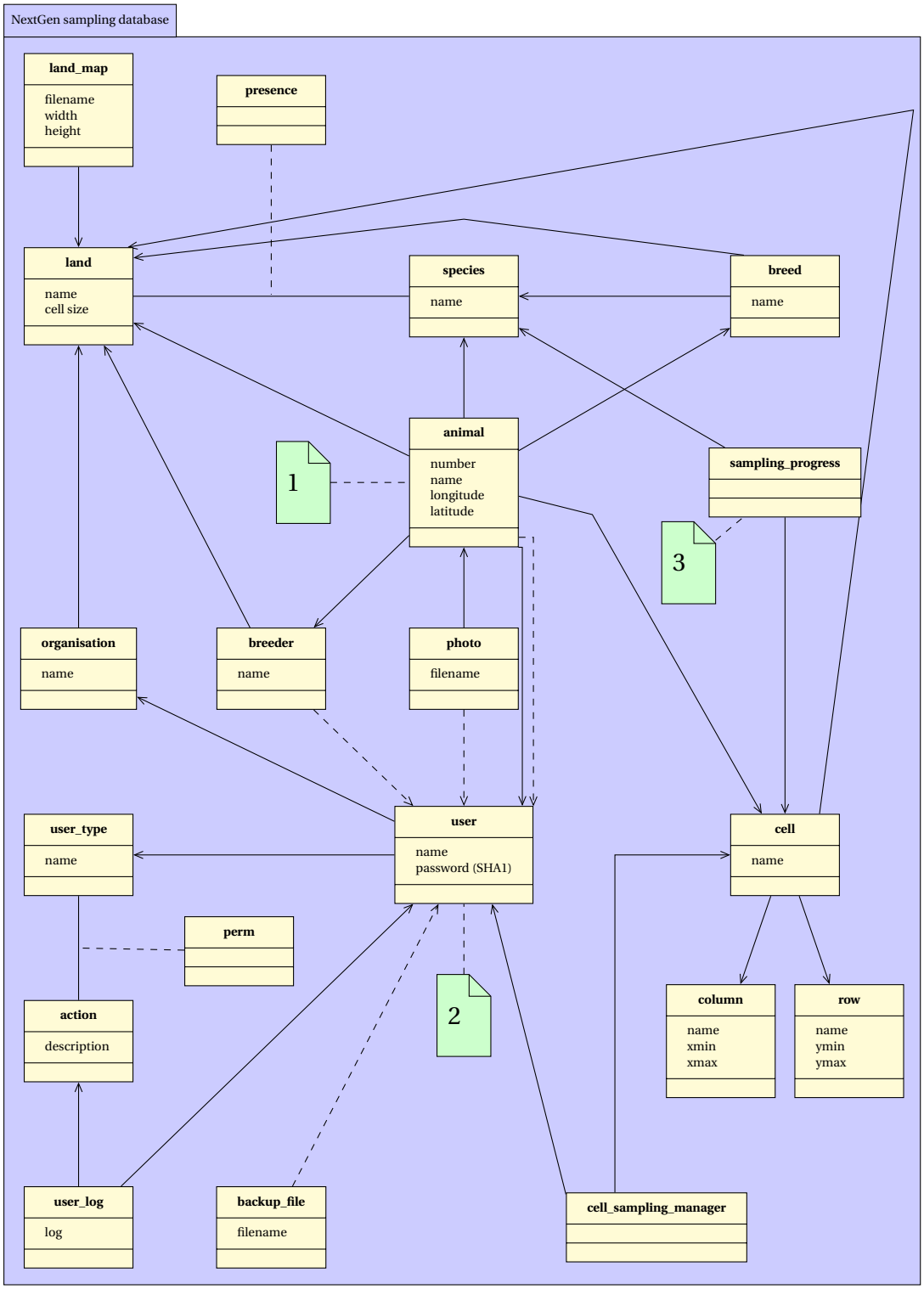
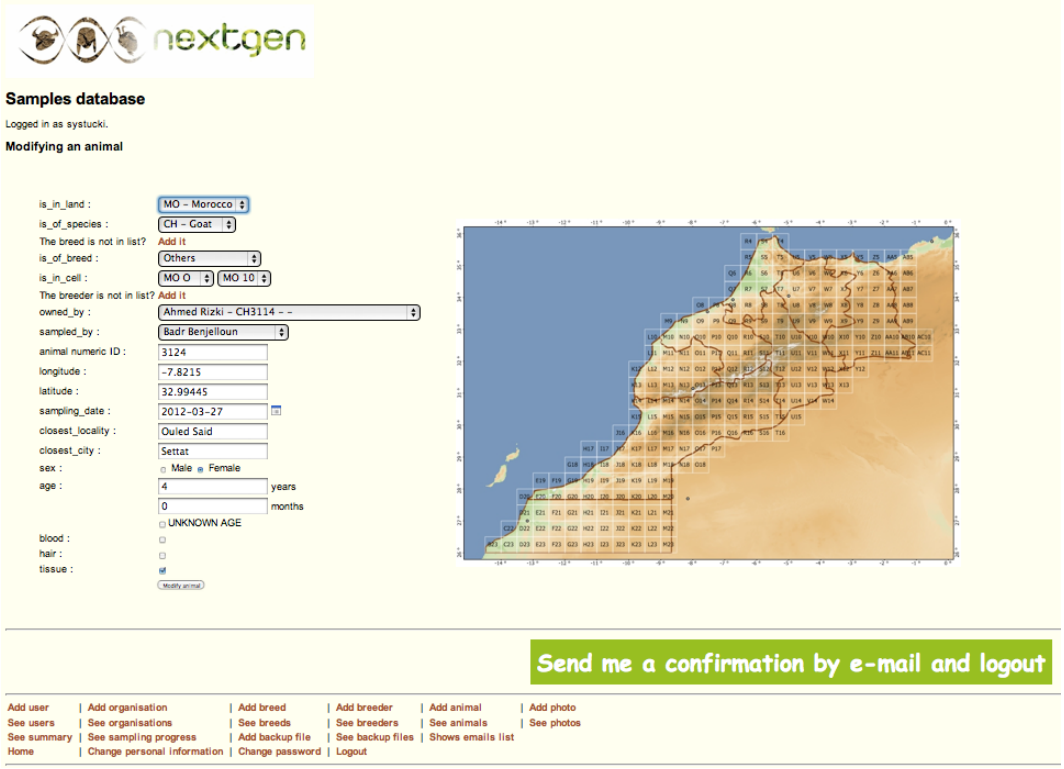
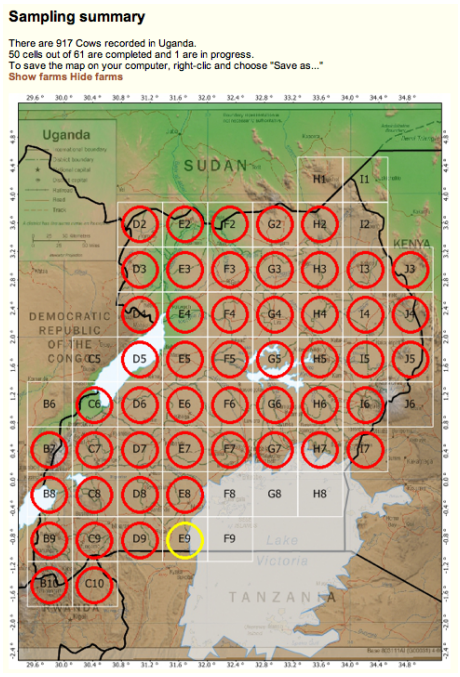


Figure A.1 – Schéma de la base de données des échantillons. Chaque élément représente une table, quelques attributs essentiels sont également indiqués. Les flèches représentent les associations entre tables, celles en traitillé enregistrent quel utilisateur a saisi un objet sur le site Web. Les tables `presence` et `perm` sont des relations multiples entre deux tables, par exemple chaque pays comprend plusieurs espèces d'animaux et chaque espèce peut être présente dans plusieurs pays. Les notes vertes se réfèrent au corps du texte.

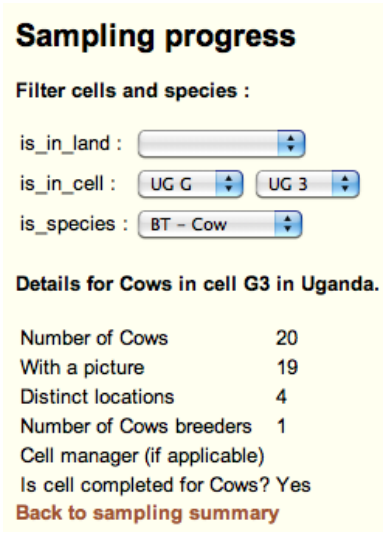


(a) La saisie et la modification d'un animal dans la base de données utilisent la même interface. La carte est un aide-mémoire pour les noms des cellules. Lorsque plusieurs animaux sont enregistrés à la suite, certaines informations comme la position et la ville la plus proche sont recopiées dans le formulaire pour faciliter la saisie.



(b) Suivi de la campagne. Les cellules où l'échantillonnage est en cours sont indiquées en jaunes et celles où il est terminé sont en rouge.

Sélection d'une cellule



(c) Détail de l'échantillonnage dans une cellule. L'utilisateur responsable d'un pays peut valider les cellules terminées depuis cet écran.

Figure A.2 – Illustrations de l'interface Web de la base de données des échantillons.

A.3 Déclaration de la base de données dans l'environnement EasyDev

La base de données sous-tendant le suivi de l'échantillonnage a été créée à partir du code source présenté dans cette section. EasyDev traite chaque élément de type « class » séparément. Chaque class est analysée pour créer, d'une part, une table dans la base de données et, d'autre part, une classe PHP qui facilite l'interaction entre le site Web et cette base. Chaque class peut contenir des attributs (par exemple des integer ou des string) et des fonctions appelées finder qui permettent de définir des requêtes SQL personnalisées (les requêtes les plus simples sont automatiquement fournies par l'environnement). Les attributs de type relation1N et relationNM servent à déclarer des relations logiques entre les tables de la base de données (ainsi que les requêtes SQL correspondantes). A la fin de la compilation, l'utilisateur obtient une base de données prête à l'emploi et un ensemble de classes PHP qu'il peut intégrer dans son site Web. C'est pourquoi chaque table de la figure A.1 correspond à une class EasyDev [2] et à une classe PHP (dont le code n'est pas reproduit dans ce document).

```
1  class land
2  {
3    relationNM species presence;
4    string name;
5    string full_name;
6    double cell_size;
7    bool sampling;
8    finder name(param_name)
9    {
10     "select * from object_land where name=param_name";
11   }
12   finder sampling()
13   {
14     "select * from object_land where sampling=TRUE";
15   }
16 }
17
18 class species
19 {
20   relationNM land presence;
21   string name;
22   string full_name;
23   bool wild;
24   finder name(param_name)
25   {
26     "select * from object_species where name=param_name";
27   }
28   finder fullname(param_fullname)
29   {
30     "select * from object_species where full_name=param_fullname";
31   }
32   finder wild(param_wild)
```

2. A l'exception des attributs de type relationNM qui sont transformés en une table de liaison entre les deux tables concernées.

```
33   {
34     "select * from object_species where wild=param_wild";
35   }
36 }
37
38 class breed
39 {
40   relation1N land is_in_land;
41   relation1N species is_of_species;
42   string name;
43   finder name(param_name)
44   {
45     "select * from object_breed where name=param_name";
46   }
47 }
48
49 class organisation
50 {
51   relation1N land is_in_land;
52   string name;
53   string full_name;
54   finder name(param_name)
55   {
56     "select * from object_organisation where name=param_name";
57   }
58   finder fullname(param_fullname)
59   {
60     "select * from object_organisation where full_name=param_fullname";
61   }
62 }
63
64 class user_type
65 {
66   relationNM action permission;
67   string name;
68   string full_name;
69   finder name(param_name)
70   {
71     "select * from object_user_type where name=param_name";
72   }
73 }
74
75 class user
76 {
77   relation1N organisation is_in_organisation;
78   relation1N user_type is_of_type;
79   string first_name;
80   string last_name;
81   string email_adr;
82   string username;
83   string password;
84   string forgotpasswordtoken;
```

```
85    finder username(param_name)
86    {
87      "select * from object_user where username=param_name";
88    }
89    finder username_password(param_name, param_password)
90    {
91      "select * from object_user where (username=param_name and password=
          param_password)";
92    }
93    finder forgotpasswordtoken(param_forgotpasswordtoken)
94    {
95      "select * from object_user where forgotpasswordtoken=
          param_forgotpasswordtoken";
96    }
97  }
98
99  class action
100 {
101   relationNM user_type permission;
102   string name;
103   string description;
104   finder name(param_name)
105   {
106     "select * from object_action where name=param_name";
107   }
108 }
109
110
111 class column
112 {
113   string name;
114   double xmin;
115   double xmax;
116   finder name(param_name)
117   {
118     "select * from object_column where name=param_name";
119   }
120 }
121
122 class row
123 {
124   string name;
125   double ymin;
126   double ymax;
127   finder name(param_name)
128   {
129     "select * from object_row where name=param_name";
130   }
131 }
132
133 class cell
134 {
```

```
135   relation1N land is_in_land;
136   relation1N column is_in_column;
137   relation1N row is_in_row;
138   string name;
139   finder name_land(param_name, param_land)
140   {
141     "select * from object_cell where (name=param_name and 1
        n_rel_is_in_land=param_land)";
142   }
143   finder bylocation(param_x, param_y)
144   {
145     "select object_cell.* from (object_cell inner join object_column on
        object_cell 1 n_rel_is_in_column=object_column.id) inner join
        object_row on object_cell 1 n_rel_is_in_row=object_row.id where (
        param_x>=object_column.xmin and param_x<object_column.xmax and
        param_y>=object_row.ymin and param_y<object_row.ymax)";
146   }
147 }
148
149 class breeder
150 {
151   relation1N land is_in_land;
152   relation1N user written_by;
153   string name;
154   string farm_ID;
155   string phone_number;
156   text address;
157   string user_ip;
158   datetime record_date;
159 }
160
161 class animal
162 {
163   relation1N land is_in_land;
164   relation1N species is_of_species;
165   relation1N breed is_of_breed;
166   relation1N cell is_in_cell;
167   relation1N breeder owned_by;
168   relation1N user sampled_by;
169   relation1N user written_by;
170   integer number;
171   string name;
172   double longitude;
173   double latitude;
174   date sampling_date;
175   string closest_locality;
176   string closest_city;
177   bool sex;
178   integer age_in_months;
179   bool blood;
180   bool hair;
181   bool tissue;
```

```
182   string user_ip;
183   datetime record_date;
184   finder name(param_name)
185   {
186   "select * from object_animal where name=param_name";
187   }
188   finder number(param_land, param_species, param_number)
189   {
190   "select * from object_animal where (1n_rel_is_in_land=param_land and 1
        n_rel_is_of_species=param_species and number=param_number)";
191   }
192 }
193
194 class photo
195 {
196   relation1N animal shows_animal;
197   relation1N user uploaded_by;
198   string filename;
199   string original_filename;
200   string user_ip;
201   datetime upload_date;
202 }
203
204 class sampling_progress
205 {
206   relation1N cell is_cell;
207   relation1N species is_species;
208   datetime record_date;
209 }
210
211 class cell_sampling_manager
212 {
213   relation1N cell is_cell;
214   relation1N user managed_by;
215   datetime record_date;
216 }
217
218
219 class backup_file
220 {
221   relation1N user uploaded_by;
222   string filename;
223   string original_filename;
224   string user_ip;
225   datetime upload_date;
226 }
227
228 class user_log
229 {
230   relation1N user is_user;
231   relation1N action is_action;
232   datetime log_time;
```

```
233   string log;
234 }
235
236 class land_map
237 {
238   relation1N land is_in_land;
239   string filename;
240   integer pic_width;
241   integer pic_height;
242   double origin_long;
243   double origin_lat;
244   integer origin_x;
245   integer origin_y;
246   double cell_width;
247   double cell_height;
248   integer mark_size;
249   integer mark_width;
250   integer point_size;
251 }
```

B Structure de Samβada

Samβada est un programme écrit en C++ dont la structure est illustrée par la figure B.1.

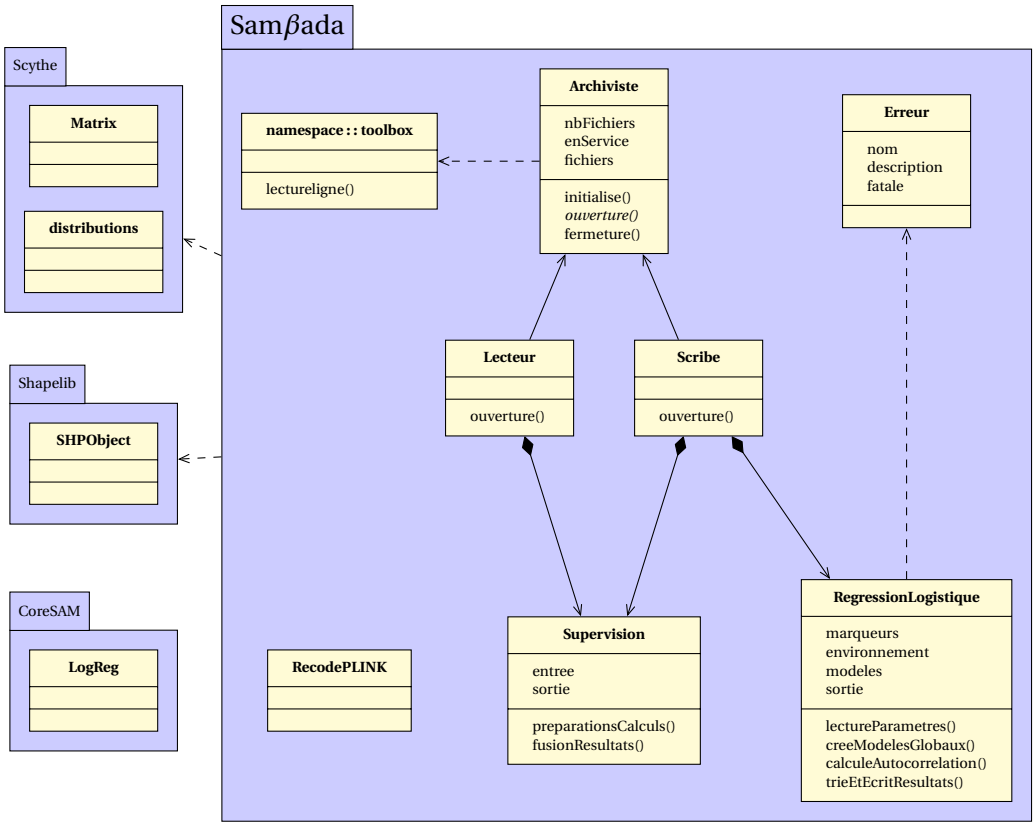


Figure B.1 – Schéma d'implémentation de Samβada. Les classes C++ sont représentées en jaune et les programmes et bibliothèques en bleu. Samβada se base sur la bibliothèque `Scythe statistical library` pour les calculs matriciels et l'estimation des distributions de probabilités. La bibliothèque `Shapelib` sert à exporter les indices locaux d'autocorrélation spatiale au format « shapefile » afin de les cartographier avec un logiciel SIG. Les principaux composants de Samβada sont les classes `Archiviste`, `Lecteur`, `Scribe` qui gèrent les entrées et sorties dans les fichiers, la classe `RegressionLogistique` qui analyse les modèles d'associations entre le génome et l'environnement et calcule l'autocorrélation spatiale et la classe `Supervision` qui s'occupe de la distribution des données et de la récolte des résultats entre plusieurs ordinateurs. Le module `RecodePLINK` permet de traduire des données moléculaires depuis le format `Plink`. `CoreSAM` est un logiciel autonome écrit en C dédié au calcul haute performance de modèles univariés.

C Article sur l'échantillonnage

La méthode de sélection des échantillons pour le séquençage a été développée en collaboration avec Diane Perez dans le cadre du cours de plans d'expériences (*Design of Experiments*) du Dr. Jean-Marie Fürbringer (cf sec. 4.4).

Nous avons pour objectif de présenter notre approche dans un article. Le rapport rédigé pour le cours doit être remanié en fonction du public de la revue choisie. Nous prévoyons *a priori* de le soumettre à *Molecular Ecology*.

Spatial Analysis to Identify Adaptive Genomic Regions: a Design of Experiment Inspired Approach

Sylvie Stucki[1], Diane Perez[2], Jean-Marie Fuerbringer[3] and Stéphane Joost[1]

[1]Lab. of Geographic information systems
[2]Solar Energy and Building Physics Lab.
[3]Doctoral school - Office of the Dean
École polytechnique fédérale de Lausanne, Switzerland

Abstract

This study concerns a collection of one thousand goats sampled over the whole territory of Morocco. Out of these, 82 to 164 subjects' genomes will be sequenced, providing for each subject information about the presence or absence of 30 millions of genetic markers. This paper proposes a method to choose the subjects to be sequenced in order to maximise the information that can be extracted about the climate's influence. To this end, the space of environmental factors to be studied is first reduced to its main axes through a Principal Component Analysis (PCA) on the values observed at each sampling location. The Euclidian distance on the main axes of the PCA is then used to perform a hierarchical ascendant clustering, allowing for the choice of samples as different as possible in terms of climatic conditions. The analysis also addresses spatial representativity of the subset. Finally, the proposed method is compared to a random selection in order to assess the validity of this approach.

1 Introduction

Understanding how plants and animals adapt to their environment is addressed by evolutionary biology. Nowadays this research fields takes on a general concern, since providing sufficient and sustainable food supply is a major challenge for our societies, especially in the context of increasing population and climate change (Food and Agriculture Organisation, 2007). While population genetics focuses on the influence of evolutionary forces on allele frequency and populations structure, landscape genetics combines spatial analysis, population genetics and molecular ecology to explain how environmental features are shaping the spatial distribution of genetic markers (Manel *et al.*, 2003). Advances in sequencing techniques made possible studies based on individuals (rather than populations) involving large molecular datasets and numerous samples. Landscape genomics was thus defined as the study of direct associations between molecular markers and environmental parameters (Luikart *et al.*, 2003).

1.1 Genetics background

Elemental genetic information is carried by *nucleotides*, the four molecules combined into long strands that entwine into chromosomes. Each cell of a living organism contains a complete set of chromosomes that form its genome. Two organisms of the same species share 99.9% of their gene pool (LaFramboise, 2009). Moreover most part of genetic variability is *neutral* because it has no effect on the fitness of an individual towards its environment. Therefore only a small fraction of the genome is subject to natural selection. The variable part of genetic data is studied based on molecular markers, accessible through sequencing

techniques. A marker is a portion of DNA defined by its position along the genome and the different versions (*alleles*) that individuals may carry.

A *Single Nucleotide Polymorphism* (SNP) is a marker consisting in a specific position (*locus*) where the nucleotide may vary among individuals so that each option is present in a significant portion of the population (LaFramboise, 2009). In this case the possible alleles are the four nucleotides (A, T, C, G). Gene pool of diploid organisms like mammals have two copies of their chromosomes, as they receive a complete set from each parent. Therefore the *genotype* of an individual at a locus includes two alleles, e. g. AA, AG or GG.

Sequencing devices decipher DNA fragments and output transcriptions (*reads*). Capillarity-based sequencers produce *reads* that are up to 700 *base pairs* (nucleotides) long. These technologies are expensive but the length of the reads ease their assembling into a genome (Gnerre *et al.*, 2011). Current high-throughput sequencers allow fast DNA processing at much lower cost than previous technologies, but they produce shorter reads ($\sim$100 bp). Processing these data requires high coverage (level of redundant reads) and specific algorithms to assemble the genome and identify variant loci (Gnerre *et al.*, 2011).

Once variant loci of a population are revealed by sequencing, genetic studies can be performed at lower cost using a device called SNP arrays which identify up to one million of predefined SNP (LaFramboise, 2009). However the range of applicability of these analysis are highly-dependent on the choice of individuals used to design them. A SNP array targeting some specific breeds may deeply underestimate the genetic variability of some other breeds that carry different SNP (*ascertainment bias*, (Morin *et al.*, 2004)).

1.2 Study context

Our study is part of the project NEXTGEN that aims at preserving farm animals biodiversity to optimise present and future breeding options. The need to improve production capacity in developing countries aroused interest in crossing traditional breeds with high-yield metropolitan ones. However offsprings may be less resilient to local conditions and livestock farming could be endangered on the long view by the extinction of the local well-adapted breeds. NEXTGEN addresses this concern by studying genetic diversity of traditional breeds using whole genome sequencing over a large panel of individuals. A part of the project is devoted to the ability of goats and sheep from Morocco to adapt to very different and sometimes extreme climatic conditions. Landscape genomics is best suited for this study since it aims to detect candidate genes for local adaptation and identify underlying natural features (Manel *et al.*, 2010; Schwartz *et al.*, 2009). Genetic material is gathered by sampling individuals in the field. The record of their locations enables for the retrieval of the environmental features of their habitats. The probability of occurrence of an allele is then computed by fitting multiple logistic regressions with the environmental parameters (Joost *et al.*, 2007). However financial constraints require an efficient sample scheme and a careful choice of the individuals for sequencing.

This study aims at selecting a collection of 164 samples meeting the following requirements:

- Climatic variability of habitats must be reliably represented to compute valid association models.
- Samples must be spread over the whole territory.
- Each breed must have enough representatives to study their fitness.

Moreover some technical requirements also apply. Using high-throughput sequencing may lead to uncertainty in SNP genotyping. Thus a set of 82 samples will be sequenced first. If data quality is sufficient for SNP calling (localisation of SNP), then other samples will be processed, otherwise the first set will be sequenced again.

A small subset of 30 samples have also to be selected from the previous one. Beside sequencing, it will be analysed using a generic goat high-density SNP array. This part of the study focuses on genetic distance between breeds, which should be comprehensively observed through the whole-genome sequencing. It intends to assess how accurately the generic SNP array measures and observes the distance between breeds.

200

1.3 Sampling

Many domains in biology require population assessment or genetic data sampling. The method used for these surveys is a recurring concern: there is no simple method to obtain a detailed data set regarding a large population on a wide area given limited means. Relying on human knowledge, intuition or inclination for most convenient method can lead to distorted results (Albert *et al.*, 2010). Therefore a regular grid was designed to perform a systematic sweep of the territory of Morocco (Hirzel and Guisan, 2002). To improve efficiency, the sampling was then targeted on farms, with three farms reviewed in each cell and three unrelated animals sampled in each farm. The regular grid ensures an access to all climatic conditions and a sample of ecological niches as large as possible.

The choice of samples to be sequenced must be made carefully in order to investigate optimally correlations of genetic data with environmental parameters. Design of experiment methods intend to rationalise the experiments performed to maximise the information provided, regardless of the domain of study. The main principle is to define precisely the factors or independent variables that could have an impact of the results of an experiment, and to chose which experiments to perform in this space of parameters in accordance with the kind of behaviour investigated (Box *et al.*, 1978). These methods assume that the factors leading to an experiment can be chosen, and that their value in their domain can be set independently, thus defining a space of parameters to explore.

Obviously, this is not the case in environmental studies, where the values of environmental factors cannot be chosen and are highly correlated. Moreover, our situation is slightly different: the sampling has already been performed in the field, and a limited number of genetic samples must be chosen amongst these to be sequenced. However, the objective is similar, namely maximising the information provided by the chosen samples.

These considerations lead to the following approach: first a principal component analysis was performed over the various meteorological conditions over the territory of Morocco. On this basis, a measure of similarity was defined as a Euclidean distance over the main directions of the factor space, limiting the weight of highly correlated factors. Then, the climatic points (i.e. the farms) were classified and grouped according to their similarity. One sample was finally chosen in each class while maximising the spatial spread, thus ensuring a regular cover of both experimental and physical spaces.

This article is organised as follows: Section 2 exposes logistic regressions and the strategy for samples choice. Results are presented in section 3. Section 4 aims at assessing the quality of the method. Concluding remarks follow in section 5.

2 Methodology

The sampling of goats on this project was performed by choosing three farms in every cell of a 0.5° regular grid over Morocco. In each farm, three animals of distinct genetic lines were sampled, producing a pool of a thousand samples. As stated above, the intent of this study is to choose three nested sample sets of 164, 82 and 30 animals for sequencing. The very limited number of genomes to be sequenced requires a judicious selection of the samples. Here we present the new methodology developed for this purpose.

2.1 Logistic regression

Landscape genomics is based on the simultaneous analysis of multiple association models between molecular markers and environmental parameters. The correlation between the occurrence of an allele and the value of a climatic factor is measured by fitting a logistic regression (Dobson and Barnett, 2008). Logit functions have their target set included between 0 and 1, thus they are suitable for modeling probabilities. If Y is a random variable representing the presence ($Y = 1$) or absence ($Y = 0$) of an allele, and x_k the value of a climatic factor X_k, we are looking for the probability of $Y = 1$ in the form of a logit function (Fig. 1) :

$$\pi(Y = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}) = \frac{e^{a+\sum_{k=1\dots P} b_k \cdot x_k}}{1 + e^{a+\sum_{k=1\dots P} b_k \cdot x_k}}$$

201

or

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = a + \mathbf{x^T}\cdot\mathbf{b}$$

with P the number of parameters, $\mathbf{X}$, $\mathbf{x}$ and $\mathbf{b}$ the vectorial notations of the variables X_k, their values x_k and the corresponding parameters a and b_k, and $\pi(\mathbf{x})$ a short notation for $\pi(Y = 1|\mathbf{X} = \mathbf{x})$.

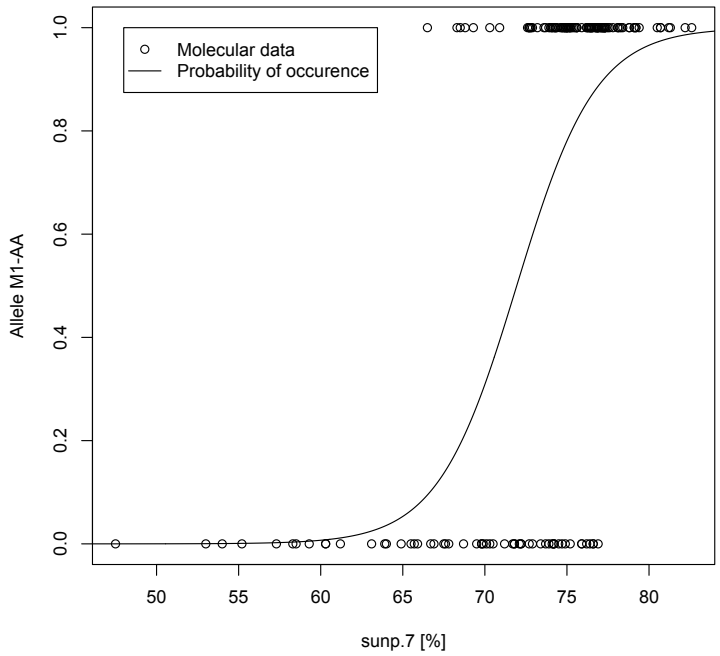


Figure 1: The logistic function modeling the probability of occurrence π of an allele subject to natural selection according to the value x_k of the environmental factor driving the adaptive process. Here the environmental factor is the sunshine in July (in percentage of daylength). The molecular data was simulated using CDPOP as explained in section 4.

The parameters a and b_k of the model are calculated by maximising the log-likelihood of all the individual experiments $i = 1 \dots N_s$.

$$l = \log L = \sum_{i=1}^{N_s} Y(i)\cdot\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) + \log\left(1-\pi(\mathbf{x}_i)\right)$$

In our case, each sequenced animal represent one experiment, with N_s the number of sequenced animals.

The significance of the model is assessed using the log-likelihood ratio (G) and the Wald statistics. The model is retained if both tests reject the null hypothesis (Joost *et al.*, 2008). The G test compares likelihoods between the current model and the null model that involves a constant only. The Wald test compares the coefficients of the model with their estimated asymptotic distributions (Dobson and Barnett, 2008).

Multivariate models include several environmental parameters to predict the repartition of the allele. Since we expect huge molecular datasets, we will first focus on univariate models only ($P = 1$). Based on our K climatic variables, we will thus consider K univariate models for each allele $j = 1 \dots J$. These models will be noted $\pi_{j,k}(x_k)$, $k = 1 \dots K$, with parameters $a_{j,k}$ and $b_{j,k}$.

2.2 Sample choice

In order to test or fit models such as the ones for the probability $\pi_{j,k}$, design of experiment methods consists in choosing the most appropriate values of the X_k factors for a limited number of experiments. In order to determine the $a_{j,k}$ and $b_{j,k}$ as precisely as possible, it is necessary to have samples corresponding to any values of the factors X_k: studying response to extreme values of the factors is not sufficient for generalised-linear models. Thus our objective will be to choose samples distributed as evenly as possible over the whole parameter space represented by the X_k. However, in our case the factors are already defined by the locations of the farms where goats were sampled.

Our study exploits temperature, pluviometry and irradiation data from the Climatic Research Unit (CRU), which extrapolates measured meteorological data over the territory (New *et al.*, 2002). The full list of environmental factors X_k is shown in table 1. It includes several probably correlated factors (such as rainfalls during each month, or rainfall and humidity data). Each climatic factor is first to be tested individually with an univariate logit function, in order to determine which factor explains best (alone) the probability of presence of an allele.

In our case, the parameter space represented by the X_k is thus discrete, but obviously highly correlated and not homogeneous.

Variable	Description
wnd	monthly values of windspeed in m/s, 10 meters above the ground
dtr	monthly values of mean diurnal temperature range in °C
frs	monthly values of number of days with ground-frost
pre	monthly values of precipitations in mm/month
pre_sigma	monthly values of the coefficient of variation of monthly precipitation in percent
tmp	monthly values of mean temperature in °C
rdo	monthly values of wet-days (number of days with >0.1 mm rain per month)
reh	monthly values of relative humidity in percent
sunp	monthly values of percent of maximum possible sunshine (percent of day length)

Table 1: Name of the topo-climatic variables, their description and abbreviation. Monthly values are available for each month, completed by a yearly mean.

2.2.1 Principal component analysis

The principal component analysis (PCA) method consists in transforming K correlated variables (in our case the environmental variables) into K or less uncorrelated ones, the *principal components* or *axes*. These new axes are obtained through a rotation of the original axes. The first principal axis is chosen to explain as much of the variance of the observations as possible. The next axes are each defined to have the highest possible variance, providing a ranking of the principal components. The axes are thus ordered according to the amount of information they support.

For our problem, the PCA will help us define what *samples as different as possible* mean: the first axes, explaining most of the divergence of our meteorological points, will be used to define a Euclidian distance between these points. This will limit the weight of highly correlated factors on our choice of samples to be sequenced. As their nature, units and values are very heterogeneous, the original factors X_k are standardised ($\tilde{X}_k$) before performing the PCA:

$$\tilde{X}_k = \frac{X_k - \bar{X}_k}{\sigma(X_k)}$$

with $\bar{X}_k$ the mean and $\sigma(X_k)$ the standard deviation of the distribution.

2.2.2 Clustering method

Once a distance is defined, a method for choosing a limited number of points representing the meteorological diversity is to regroup the points (farms) according to their climatic closeness and to choose one in each cluster. The N points are classified using the hierarchical agglomerative clustering method with the Ward's criterion. This clustering process starts with N clusters containing one point and then iteratively merge the closest ones to obtain a clustering of N-1 groups. This process is deterministic and builds a hierarchical tree of N nested clusterings composed of 1 to N clusters. The tree was cut at the 164- and 82-clusters levels and one sample was chosen in each cluster. The approach ensures an even representation of climatic conditions.

2.3 Constrained sample drawing

The sampling scheme based on a regular grid provided an even density of samples over the territory. However choosing a sample at random in each climatic cluster may reintroduce spatial bias. The sprawl of selected samples over the territory was assessed by a spatial clustering index. This measure of the global spread of a sample set was defined as the sum of the distances between each point and its nearest neighbour in the set. The higher the index, the farther the distance between neighbours and thus the better the sprawl over the territory.

3 Results

3.1 Principal component analysis

The PCA performed on the K=117 climatic factors confirmed their highly correlated nature, with 96% of the variance explained by the first 7 principal axes. The main correlations of the two first axes with the original climatic factors are shown in Table 2. Principal component 1 and 2 account respectively for 44% and 30% of the total variance.

The main axes are not linked with a particular environmental variable, however some trends appear. The first principal axis shows positive correlations with the amount of rainy days and negative ones with sunshine and variation of precipitation. As illustrated on Fig. 2 axis 1 differentiates coastal and deserts regions of Morroco. The second principal component is associated with high diurnal temperature range, many days with ground-frost, high amount precipitation and low temperature during winter. Thus Fig. 3 shows that axis 2 is correlated with altitude.

The space generated by these 7 axes represents reliably our climatic conditions, allowing for the use of the Euclidian distance in this subspace for the clustering of the farms.

3.2 Clustering

Of the original $N_f = 412$ farms where samples were collected, 226 actually have different climatic conditions. Farms with identical environmental data define a unique *climatic point* in the space of environmental parameters. The classification was thus performed on the $N_c = 226$ climatic points. The grouping of the farms (through the corresponding climatic points) at the 10-cluster level of the tree is shown in Fig. 4.

The clusterings of interest for our study are the 82- and 164-cluster levels. The location of the 164 clusters relatively to several original climatic factors is shown in Fig 5. As expected, each class covers a limited range of values for each climatic variable.

3.3 Constrained sample drawing

At the 164-classes level of the classification, classes contain 1 to 3 farms and 1 to 18 samples to choose from. For representation issues, three individuals of each known race with very few samples were first chosen in different classes. Then the sample of one animal per remaining

Principal component 1		Principal component 2	
Name	Projection	Name	Projection
pre_sigma.1	-0.12588	dtr.7	0.13723
pre_sigma.2	-0.12941	dtr.8	0.13855
pre_sigma.3	-0.12600	frs.1	0.15954
pre.2	0.12625	frs.11	0.14393
pre.3	0.12429	frs.12	0.15803
rd0.1	0.12977	frs.2	0.15402
rd0.10	0.12357	frs.3	0.14827
rd0.11	0.12856	frs.4	0.13886
rd0.12	0.12956	frs.avg	0.15593
rd0.2	0.13118	pre.7	0.14041
rd0.3	0.13201	pre.8	0.15722
rd0.4	0.12867	pre.9	0.15022
rd0.avg	0.12656	rd0.7	0.15294
reh.4	0.12359	rd0.8	0.16157
sunp.1	-0.12485	rd0.9	0.16062
sunp.11	-0.13004	tmp.1	-0.16573
sunp.12	-0.12688	tmp.11	-0.15915
sunp.2	-0.12813	tmp.12	-0.16638
sunp.3	-0.12829	tmp.2	-0.15598
sunp.4	-0.12875	tmp.3	-0.13678

Table 2: Projections of environmental variables on the two first principal axes. The 20 highest correlations are sorted by environmental groups.

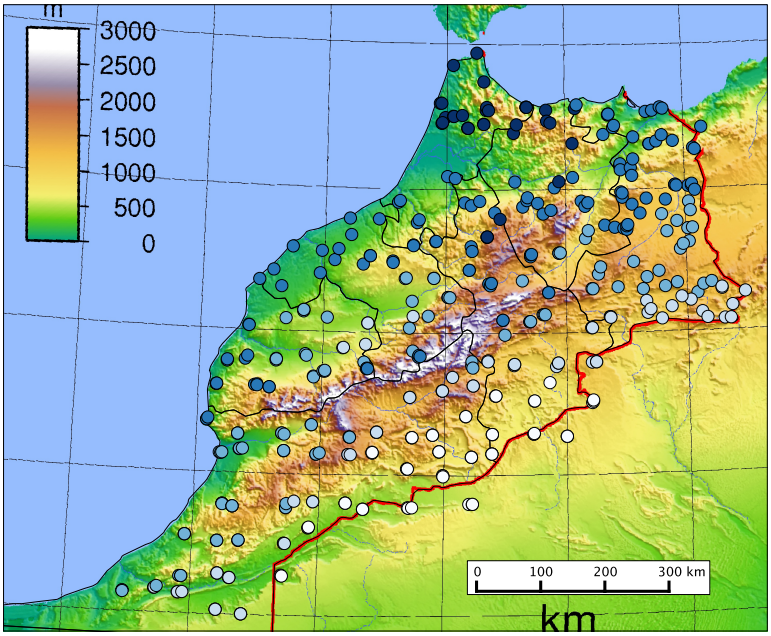


Figure 2: Principal component 1 differentiates coastal and desert regions of Morroco. High values (in blue) are correlated with the number of rainy days while low values (in white) are found near the desert where sunshine is high and variation of precipitation is low. Source of the altitude data: Wikipedia.

205

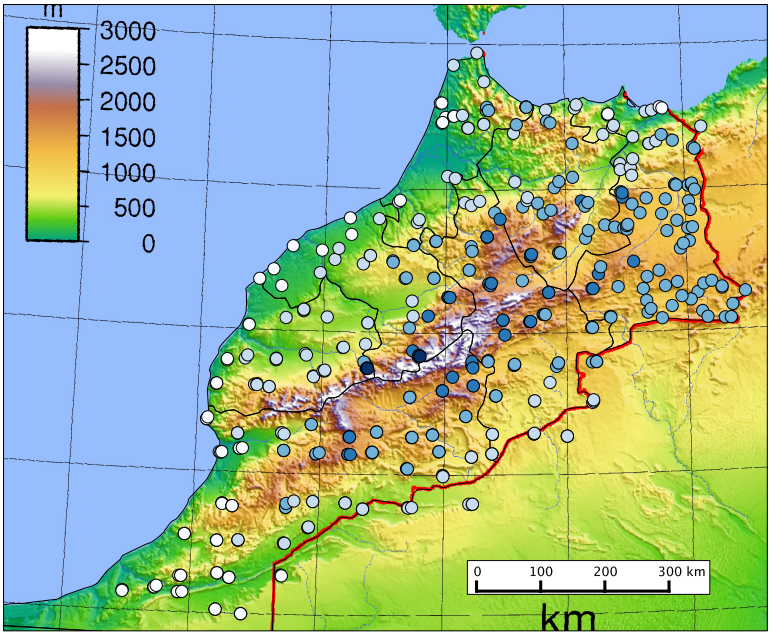


Figure 3: Principal component 2 is correlated with altitude. High values (in blue) are found in the Atlas mountains where winter temperature is low while diurnal temperature range, number of days with frost ground and precipitation are high. Source of the altitude data: Wikipedia.

class was added to the selection. The 164 samples are thus chosen in different farms representing the 164 classes. Several candidate sets were drawn in order to maximise the spatial spread.

The 82 samples to be sequenced first were chosen amongst the 164 using this time the classification at the 82-classes level. At this level, the 164 samples were regrouped in 82 classes of 1 to 4 elements. One sample was again chosen in each class, while maximising the total distance between the samples and guarantying breed representation.

The subset of 30 individual is intended for the SNP array genotyping. The former approach could not be applied to select this 30-sample subset, due to the requirement of breed representativity which could not be met using the 30-group clustering. Thus "mini-sets" of 30 samples with an appropriate breed repartition were drawn from the 82-sample subset and the candidate with the highest spatial index was chosen.

The locations of the farms for the 30, 82 and 164 samples selections are shown in Fig 6. Fig 7 illustrates how the 164 selected farms are distributed on the 7 principal axes of the PCA analysis. It shows that the farms are already well distributed along the principal axes, thus limiting the regulatory role of the clustering method. A lucky random selection might have given a similar picture, but without any guaranty of representativeness.

4 Proof of concept

4.1 Method

The search of an appropriate subset of samples for sequencing raises two questions about its size and its efficiency. Is our subset large enough to detect adaptive mutations that would appear if all samples could be analysed? How does our subset perform compared to a random subset of the same size? We need molecular data to study these matters. Since actual data will not be available to a large extent of the samples, we will use simulated molecular data for every animal.

CDPOP is a software for simulating the evolution of a population on an individual basis (Landguth and Cushman, 2010). Each individual is depicted by its location, age and

206

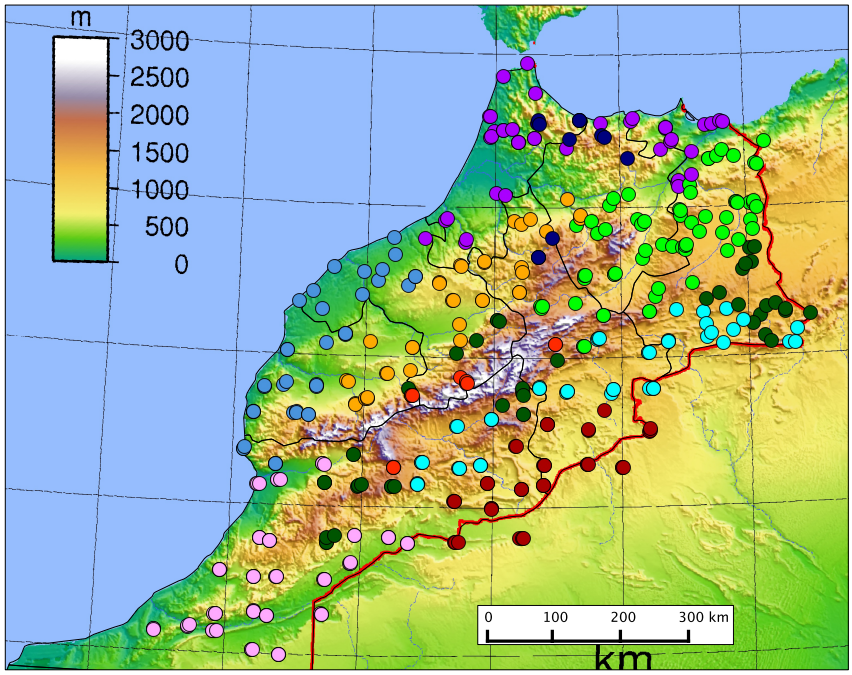


Figure 4: Locations of the farms classified in 10 categories (different colors). As could be expected, the climatically similarity is partly correlated to the spatial proximity and similarity in altitude. The large scale geography also dissociates coastal and inland areas. Source of the altitude data: Wikipedia.

genotype. The locations are fixed, they can be empty or occupied by an individual. At each iteration the population mates and the offspring migrate to possibly replace another individual. Both events can be spatially random or occur with a probability depending on the distance between locations. The genotype of newborns is drawn according to their parents and the mutation rate. A typical run can simulate a population of one thousand individuals with one hundred of neutral loci over one thousand generations. Current version of CDPOP enables the inclusion of one or two loci subject to natural selection. The fitness of an individual is modelised by a specific birth mortality rate depending on the newborn's genotype and the location (Landguth *et al.*, 2012).

A goat population whose possible locations were the actual positions of our samples was simulated. Up to three individuals were allowed in each farm. The inverse of the Euclidean distance was used as the cost function for mating and migration. Genetic data consisted in five loci (M1 to M5) having three possible alleles AA, AG and GG. The M1 loci was subject to natural selection under the influence of the duration of sunshine in July (sunp.7). The range of the environmental factor was divided in three parts as shown on Fig 8a. The genotype AA was adapted to high values, AG to middle-ranged values and GG to low values of sunshine. Two sets of fitness surfaces were built for the simulation. The first scenario had a constant birth mortality rate for each environmental sub-range. The second one had a varying rate to obtain a continuous fitness function. Fig. 8b-d presents the resulting birth mortality rates for each genotype with both scenarios. Starting with random genotypes, we let the goat population evolve over 1000 generations in each case. Our molecular data consisted in the genotypes of the final population. Since CDPOP keep locations constant, each actual sample received the genotype of the corresponding simulated individual. Genetic data was recoded to account for the presence or the absence of each possible allele M1-AA, M1-AG, . . . , M5-GG, thus a total of 15 genetic markers were considered.

Based on this simulated genetic data, all the models defined in section 2.1 were tested, against five sets of animals :

1. All 1117 animals

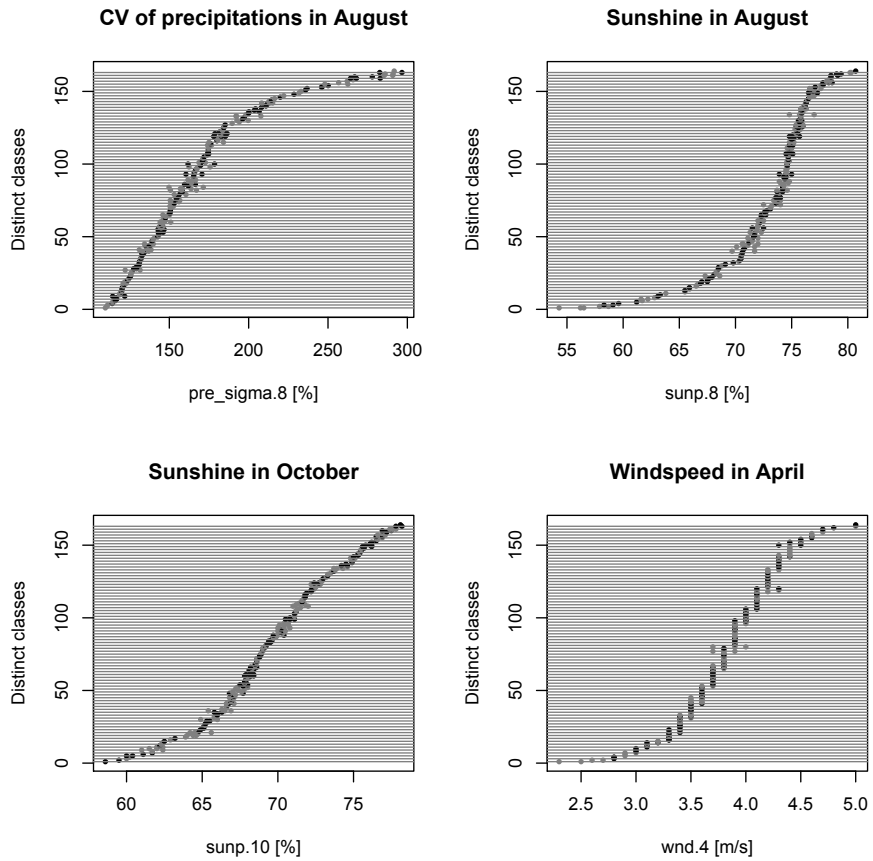


Figure 5: The 164 clusters are shown here relatively to four original environmental variables. Each line represents a cluster, with the value taken by its climatic points for the current variable. To improve the discernability, the clusters have been ordered according to the mean value of their points.

2. The 164 selected animals
3. A random 164-animal selection
4. The 82-animal sub-selection
5. A random 82-animal sub-selection of the set 3.

4.2 Results

The first results of this study are not clear enough for final conclusions and call for further investigations. However, some interesting preliminary remarks can be made. Table 3 shows a summary of our models. Focus is made on the ability to identify neutral and adaptive loci, thus the three markers accounting for a loci are considered as a whole.

The model (M1, sunp.7) was successfully detected by both selections of 164 samples, while smaller sets failed. The threshold value for model significance does not depend on the number of regression points, while the G and Wald scores do, hence more models are detected by the complete set of animals. The performance of the two sets of 82 samples are similarly poor. However, our selection detect less spurious correlations with other markers than the random 82-animals set. It can be noted that this last set is also based on our regular grid scheme, and thus can be expected to perform better than an entirely unplanned sampling.

In the four first cases, our model of interest is among the 20 best associations. Strikingly, the environmental factor sunp.7 never appears in the best model for M1. It is however

208

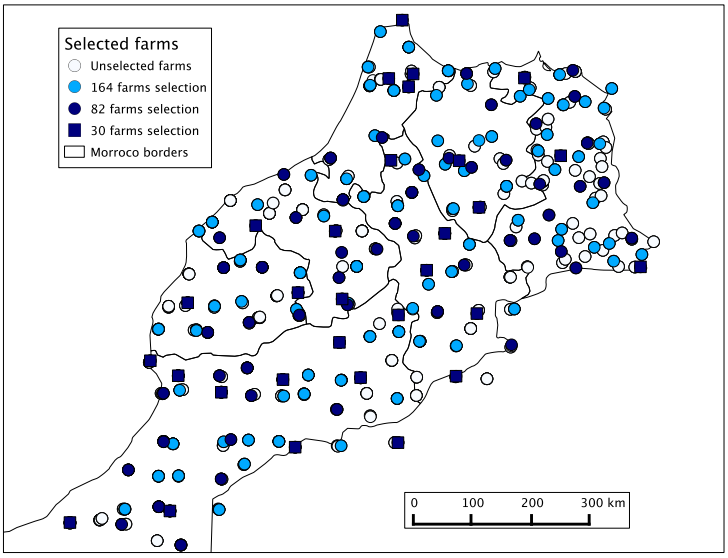


Figure 6: Locations of the selected animals' farms. The 82 sample selection is included in the 164 sample selection. The subset of 30 samples were chosen among the 82 samples for complementary SNP genotyping.

Variable	All samples	164-selection	164-random	82-selection	82-random
Model (M1, sunp.7) accepted	yes	yes	yes	no	no
Total accepted models	830	110	107	3	0
Model (M1, sunp.7) position	5	14	3	15	997
M1 models in top 20	17	20	16	12	12
Best M1 model with:	pre_sigma.9	rd0.5	pre_sigma.9	rd0.5	pre_sigma.4
(correlation with sunp.7)	(-0.86)	(0.51)	(-0.86)	(0.51)	(-0.6)
Position first other model	11	21	13	1	4
First other model	M3 sunp.7	M3 sunp.7	M3 sunp.7	M3 sunp.7	M3 pre_sigma.7

Table 3: Main results of the proof-of-concept study, based on the simulated data obtained with the continuous birth mortality rate model. Models are named after the loci and the environmental variable involved. The models were ordered according to the G-score. Threshold values for model acceptance: 20.58 for both Wald and G-score.

correlated with the best explaining factors, suggesting that with highly correlated variables and a limited number of sample, singling out a *best* explaining factor is probably overreached.

Some models with loci M3 were accepted, although M3 is supposed to be neutral. M3 is thus involved in the first "spurious" model in each case, four times in association with sunp.7. We can remark that M1 and M3 are correlated in the whole dataset, σ(M1-AA, M3-AA) = 0.46, σ(M1-GG, M3-GG) = 0.59, thus explaining the observed associations between M3 and the environmental factors. Real genetic markers which are close on the genome are often transmitted together to offsprings. This is known as *linkage disequilibrium* and such markers are correlated. However CDPOP documentation does not report this to be accounted for in the simulations. Some further investigations are needed to understand why the simulation designed M3 this way.

The main issue in this experiment is the low significance of models based on 82 animals. If the simulated data are similar to natural ones, all 164 samples will be needed to build reliable models. However our lack of experience with CDPOP calls for further investigations to ensure the simulation was correct. Among the points that might have hindered the simulation is the constraint on location (the actual sampling farms only). To balance this limitation,

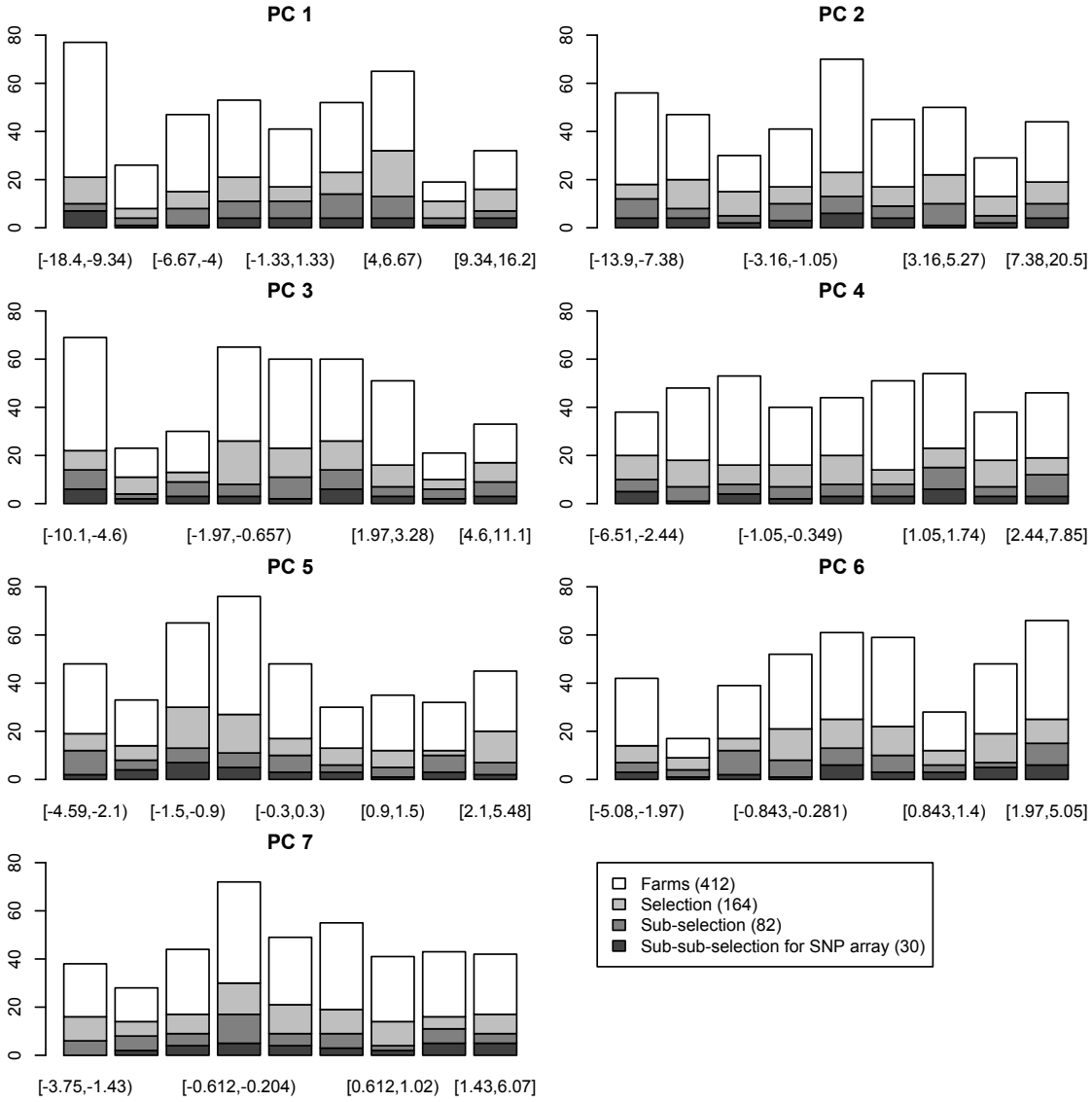


Figure 7: Distribution of the selected farms on the main axes. The three sets of selected samples are nested.

individuals were allowed to mate and move on longer distances than they usually cover in reality. Secondly the fertility was set high enough to obtain one animal per location at the end of simulation. At each generation, the animals which could not be attributed to an available location were removed, which might also have introduce some bias.

5 Conclusion

This paper is part of an ongoing study on the correlation between climatic conditions and the genetic material of goats over the territory of Morocco. It proposes a method to select a limited number of genetic samples to be sequenced among a larger pool of collected samples, in order to maximise the information that can be extracted about the correlation between climatic factors and allele's presence. Indeed, the sampling method, here a spatially distributed sweep, does not exclude the over-representation of particular climatic conditions which could obscure the dynamic over the whole range of existing conditions. Thus the selection of a limited number of sample for a detailed study must be performed carefully.

The method discussed here consists first in a study of the climatic conditions space, using PCA, in order to determine the principal components and define a "climatic" distance

210

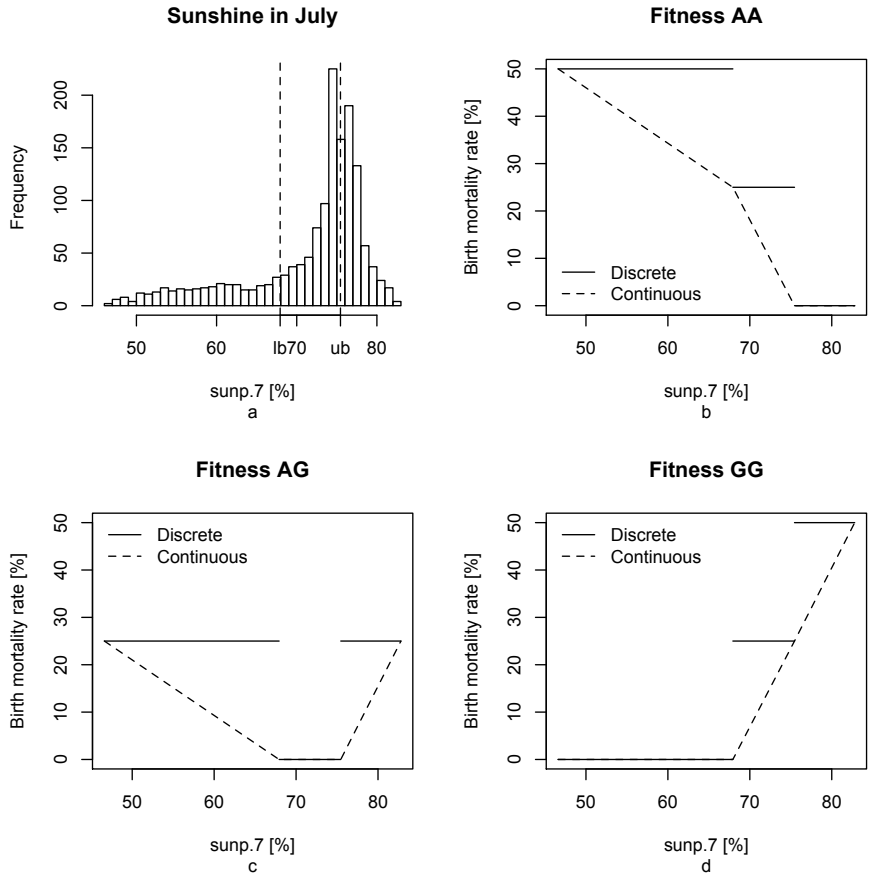


Figure 8: *a*) Distribution of the environmental condition "duration of sunshine in July" (sunp.7). Its range was cut in three parts to allow distinct fitness responses. The bounds are the mean plus or minus half the standard deviation. *b*) Birth mortality rates for individuals with genotype AA. Their fitness is higher in locations with high values of sunp.7. *c*) Individuals with genotype AG have higher fitness for middle-ranged values of sunp.7. *d*) Individuals with genotype GG are more adapted to low values of sunp.7.

between sampling locations. Secondly, the hierarchical agglomerative clustering method is used to regroup the closest locations in a number of clusters corresponding to the number of samples to select. Finally, randomly choosing one sample in each cluster guarantees a good representation of all climatic conditions.

In our case, the distribution of the sampling locations in the climatic factor space happened to be quite good, limiting the potential benefits of the clustering method to choose the samples to be sequenced. Nevertheless, this method avoids the risk of a badly distributed random choice.

An attempt to assess the validity of our approach was made with data simulated by CDPOP. The evolution of five loci were simulated, with one subject to selection under a climatic variable. Our 164-sample set correctly detects the loci subject to selection, but finds higher correlations with other climatic variables. The 82-sample set proved insufficient to build significant models, which underlines a possible issue if the second half of samples are not sequenced. However uncertainties remain about the capacity of our simulation to imitate actual genetic data, and more analysis will be performed to confirm and detail the first deductions discussed here.

References

Albert, C. H., Yoccoz, N. G., Edwards, Jr., T. C., Graham, C. H., Zimmermann, N. E., and Thuiller, W. (2010). Sampling in ecology and evolution - bridging the gap between theory and practice. *Ecography*, **33**(6, SI), 1028–1037.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons, New York, 1st edition.

Dobson, A. J. and Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall, 3rd edition.

Food and Agriculture Organisation (2007). Global plan of action for animal genetic resources and the Interlaken declaration. Rome.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(4), 1513–1518.

Hirzel, A. and Guisan, A. (2002). Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**(2-3), 331–341.

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G., and Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**(18), 3955–3969.

Joost, S., Kalbermatten, M., and Bonin, A. (2008). Spatial Analysis Method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Ressources*, **8**, 957–960.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, **37**(13), 4181–4193.

Landguth, E. L. and Cushman, S. A. (2010). CDPOP: A spatially explicit cost distance population genetics program. *Molecular Ecology Ressources*, **10**(1), 156–161.

Landguth, E. L., Cushman, S. A., and Johnson, N. A. (2012). Simulating natural selection in landscape genetics. *Molecular Ecology Ressources*, **12**(2), 363–368.

212

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power of population genomics: from genotyping to genome typing. *Trends in Ecology and Evolution*, **4**(12), 981–994.

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution*, **18**(4), 189–197.

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S., Scribner, K. T., Bonin, A., and Fortin, M.-J. (2010). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**(17), 3760–3772.

Morin, P. A., Luikart, G., Wayne, R. K., and Grp, S. W. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution*, **19**(4), 208–216.

New, M., Lister, D., Hulme, M., and Makin, I. (2002). A high-resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Schwartz, M. K., Luikart, G., McKelvey, K. S., and Cushman, S. A. (2009). Landscape genomics: A brief perspective. In Springer, editor, *Spatial Complexity, Informatics and Wildlife Conservation*, chapter 9, pages 165–174. Cushman, Samuel A. and F. Huettmann.

Bibliographie

Agha, Saif (2011), *Landscape genomics and F_{ST} approaches to reveal SNPs under selection in Cattle*, rapp. tech., European Science Foundation.

Ajmone Marsan, Paolo, Jose Fernando Garcia, Johannes A. Lenstra et Globaldiv Consortium (2010), « On the Origin of Cattle : How Aurochs Became Cattle and Colonized the World », *Evolutionary Anthropology* **19**(4), 148–157.

Albert, Cécile H., Nigel G. Yoccoz, Thomas C. Edwards Jr. et al. (2010), « Sampling in ecology and evolution - bridging the gap between theory and practice », *Ecography* **33**(6, SI), 1028–1037.

Alexander, David H., John Novembre et Kenneth Lange (2009), « Fast model-based estimation of ancestry in unrelated individuals », *Genome Research* **19**(9), 1655–1664.

– (2013), *Admixture 1.23 Software Manual*, University of California Los Angeles.

Andrew, Rose L., Louis Bernatchez, Aurélie Bonin et al. (2013), « A road map for molecular ecology », *Molecular Ecology* **22**(10), 2605–2626.

Anselin, Luc (1995), « Local Indicators of Spatial Association - LISA », *Geographical Analysis* **27**(2), GISDATA (Geographic Information Systems Data) Specialist Meeting on GIS (Geographic Information Systems) and Spatial Analysis, Amsterdam, Netherlands, Dec 01-05, 1993, 93–115.

Anselin, Luc, Ibnu Syabri et Oleg Smirnov (2002), « Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows », in : *New Tools for Spatial Data Analysis : Proceedings of the Specialist Meeting*, sous la dir. d'University of California Center for Spatially Integrated Social Science (CSISS), Santa Barbara, CA : Luc Anselin et Sergio Rey.

Antao, T. et Mark A. Beaumont (2011), « Mcheza : a workbench to detect selection using dominant markers », *Bioinformatics* **27**(12), 1717–1718.

Antao, Tiago, Ana Lopes, Ricardo Lopes, Albano Beja-Pereira et Gordon Luikart (2008), « LOSITAN : A workbench to detect molecular adaptation based on a F_{ST}-outlier method », *BMC Bioinformatics* **9**(1), 323.

Beaumont, Mark A. et David J. Balding (2004), « Identifying adaptive genetic divergence among populations from genome scans », *Molecular Ecology* **13**(4), 969–980.

Beaumont, Mark A. et R. A. Nichols (1996), « Evaluating loci for use in the genetic analysis of population structure », *Proceedings Royal Society London B* **263**, 1619–1626.

Benjamini, Y et Y Hochberg (1995), « Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing », *Journal of the Royal Statistical Society Series B-Statistical Methodology* **57**(1), 289–300.

Benson, John F., Brent R. Patterson et Tyler J. Wheeldon (2012), « Spatial genetic and morphologic structure of wolves and coyotes in relation to environmental heterogeneity in a Canis hybrid zone », *Molecular Ecology* **21**(24), 5934–5954.

Berthouly-Salazar, Cecile, Sophie Thevenon, Thu Nhu Van et al. (2012), « Uncontrolled admixture and loss of genetic diversity in a local Vietnamese pig breed », *Ecology and Evolution* **2**(5), 962–975.

Bett, R. C., M. G. Gicheha, I. S. Kosgey, A. K. Kahi et K. J. Peters (2012), « Economic values for disease resistance traits in dairy goat production systems in Kenya », *Small Ruminant Research* **102**(2-3), 135–141.

Biggar, Paul, Edsko de Vries et David Gregg (2012), « A practical solution for achieving language compatibility in scripting language compilers », *Science of Computer Programming* **77**(9), <ce :title>The Programming Languages track at the 24th {ACM} Symposium on Applied Computing (SAC'09)</ce :title>, 971–989.

Bolker, Benjamin M., Mollie E. Brooks, Shane W. Clark Connie J. asannd Geange et al. (2009), « Generalized linear mixed models : a practical guide for ecology and evolution », *Trends in Ecology and Evolution* **24**(3), 127–135.

Bonferroni, C. E. (1936), « Teoria statistica delle classi e calcolo delle probabilità », *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62.

Borcard, D et P Legendre (2002), « All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices », *Ecological Modelling* **153**(1-2), 51–68.

Bothwell, Helen, Sarah Bisbing, Nina Overgaard Therkildsen et al. (2013), « Identifying genetic signatures of selection in a non-model species, alpine gentian (Gentiana nivalis L.), using a landscape genetic approach », *Conservation Genetics* **14**(2, SI), 467–481.

Box, George E. P., William G. Hunter et J. Stuart Hunter (1978), *Statistics for Experimenters : An Introduction to Design, Data Analysis, and Model Building*, 1re éd., New York : John Wiley & Sons.

Bruford, Michael W. (2011), *Wildlife biodiversity and conservation management : are there any lessons to be learned ?*, EPFL, Lausanne, URL : http://www.globaldiv.eu/Lausanne/Bruford.pdf.

Butler, Jonathan, Iain MacCallum, Michael Kleber et al. (2008), « ALLPATHS : De novo assembly of whole-genome shotgun microreads », *Genome Research* **18**(5), 810–820.

Carl, G. et I. Kuehn (2007), « Analyzing spatial autocorrelation in species distributions using Gaussian and logit models », *Ecological Modelling* **207**(2-4), 159–170.

Cavalli-Sforza, L. L. (1966), « Population Structure and Human Evolution », *Proceedings of the Royal Society B-Biological Sciences* **164**(995), 362–379.

Chalmers, J. N. M., Elizabeth W. Ikin et A. E. Mourant (1953), « A Study Of Two Unusual Blood-Group Antigens In West Africans », *The British Medical Journal* **2**(4829), 175–177.

Coop, Graham, David Witonsky, Anna Di Rienzo et Jonathan K. Pritchard (2010), « Using Environmental Correlations to Identify Loci Underlying Local Adaptation », *Genetics* **185**(4), 1411–1423.

Coulon, A, G Guillot, JF Cosson et al. (2006), « Genetic structure is influenced by landscape features : empirical evidence from a roe deer population », *Molecular Ecology* **15**(6), 1669–1679.

Crow, James F. (2008), « Mid-Century Controversies in Population Genetics », *Annual Review of Genetics*, Annual Review of Genetics **42**, 1–16.

Darwin, Charles (1859), *On the Origin of Species*, London : John Murray.

De Mita, Stéphane, Anne-Céline Thuillet, Laurène Gay et al. (2013), « Detecting selection along environmental gradients : analysis of eight methods and their effectiveness for outbreeding and selfing populations », *Molecular Ecology* **22**(5), 1383–1399.

Dobson, Annette J. et Adrian G. Barnett (2008), *An Introduction to Generalized Linear Models*, 3^{e} éd., Chapman & Hall.

Dray, S., P. Legendre et P.R. Peres-Neto (2006), « Spatial modelling : a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM) », *Ecological Modelling* **196**(3-4), cited By (since 1996)238, 483–493.

Emery, Pauline (2012), « GEOME, a web-based platform for an integrated analysis of spatial environmental and molecular data », mém.de mast., École polytechnique fédérale de Lausanne.

Engle, Robert F. (1983), « Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics », in : *Handbook of Econometrics II*, sous la dir. de M. D. Intriligator et Z. Griliches, Elsevier, 796–801.

Epperson, Bryan K., Brad H. McRae, Kim Scribner et al. (2010), « Utility of computer simulations in landscape genetics », *Molecular Ecology* **19**(17, SI), 3549–3564.

Escoffier, Brigitte et Jérôme Pagès (2008), *Analyses Factorielles Simples et Multiples*, 4^{e} éd., Paris : Dunod.

Excoffier, Laurent et Gerald Heckel (2006), « Computer programs for population genetics data analysis : a survival guide », *Nature Reviews Genetics* **7**(10), 745–758.

Excoffier, Laurent, T. Hofer et Matthieu Foll (2009), « Detecting loci under selection in a hierarchically structured population », *Heredity* **103**(4), 285–298.

Excoffier, Laurent et Heidi Lischer (2011), *Arlequin : An Integrated Software Package for Population Genetics Data Analysis (ver 3.5)*, University of Bern.

Excoffier, Laurent et Heidi E. L. Lischer (2010), « Arlequin suite ver 3.5 : a new series of programs to perform population genetics analyses under Linux and Windows », *Molecular Ecology Ressources* **10**(3), 564–567.

Excoffier, Laurent, Pe Smouse et Jm Quattro (1992), « Analysis of molecular variance inferred from metric distances among DNA haplotypes - Application to human mitochondrial-DNA restriction data », *Genetics* **131**(2), 479–491.

Farr, Tom G., Paul A. Rosen, Edward Caro et al. (2007), « The shuttle radar topography mission », *Reviews of Geophysics* **45**(2).

Flemming, Walther (1882), *Zellsubstanz, Kern und Zelltheilung*, Leipzig : F. C. W. Vogel.

Flicek, Paul, Ikhlak Ahmed, M. Ridwan Amode et al. (2013), « Ensembl 2013 », *Nucleic Acids Research* **41**(D1), D48–D55, eprint : http://nar.oxfordjournals.org/content/41/D1/D48.full.pdf+html.

Foll, Matthieu et Oscar Gaggiotti (2008), « A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers : A Bayesian Perspective », *Genetics* **180**, 977–993.

Fotheringham, A. Stewart, Chris Brunsdon et Martin Charlton (2002), *Geographically Weighted Regression : the analysis of spatially varying relationships*, 1re éd., Chichester : John Wiley & Sons.

Frichot, Eric, Sean D. Schoville, Guillaume Bouchard et Olivier François (2013), « Testing for associations between loci and environmental gradients using latent factor mixed models », *Molecular Biology and Evolution* **30**(7), 1687–1699, eprint : http://mbe.oxfordjournals.org/content/early/2013/03/29/molbev.mst063.full.pdf+html.

Gepts, Paul et Roberto Papa (2003), « Evolution during domestication », in : *Encyclopedia of Life Sciences*, Chichester : John Wiley & Sons, Ltd.

Gillespie, JH (2001), « Is the population size of a species relevant to its evolution ? », *Evolution* **55**(11), 2161–2169.

Gillespie, John H. (2000), « Genetic drift in an infinite population : The pseudohitchhiking model », *Genetics* **155**(2), 909–919.

Gnerre, Sante, Iain MacCallum, Dariusz Przybylski et al. (2011), « High-quality draft assemblies of mammalian genomes from massively parallel sequence data », *Proceedings of the National Academy of Sciences of the United States of America* **108**(4), 1513–1518.

Goddard, Michael E., Ben J. Hayes et Theo H. E. Meuwissen (2010), « Genomic selection in livestock populations », *Genetics Research* **92**(5-6), 413–421.

Groeneveld, L. F., J. A. Lenstra, H. Eding et al. (2010), « Genetic diversity in farm animals - a review », *Animal Genetics* **41**(1), 6–31.

Guillot, Gilles, Renaud Vitalis, Arnaud le Rouzic et Mathieu Gautier (2013), « Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies », *Spatial Statistics* **(in press)**, pages.

Guisan, A et NE Zimmermann (2000), « Predictive habitat distribution models in ecology », *Ecological Modelling* **135**(2-3), 147–186.

Gunderson, Kevin L., Frank J. Steemers, Grace Lee, Leo G. Mendoza et Mark S. Chee (2005), « A genome-wide scalable SNP genotyping assay using microarray technology », *Nature Genetics* **37**(5), 549–554.

Günther, Torsten et Graham Coop (2013), « Robust Identification of Local Adaptation from Allele Frequencies », *Genetics* **195**(1), 205–220, eprint : http://www.genetics.org/content/195/1/205.full.pdf+html.

Hanotte, Olivier, Tadelle Dessie et Steve Kemp (2010), « Time to Tap Africa's Livestock Genomes », *Science* **328**, 1640–1641.

Hardy, Godfrey H. (1908), « Mendelian proportions in a mixed population », *Science* **28**, 49–50.

Harris, H (1966), « Enzyme Polymorphisms In Man », *Proceedings of the Royal Society B-Biological Sciences* **164**(995), 298–&.

Hastie, Trevor et Robert Tibshirani (1986), « Generalized Additive Models », *Statistical Science* **1**(3), 297–310.

Henry, Jean-Pierre et Pierre-Henri Gouyon (2008), *Précis de génétique des populations*, Paris : Dunod.

Hijmans, RJ, SE Cameron, JL Parra, PG Jones et A Jarvis (2005), « Very high resolution interpolated climate surfaces for global land areas », *International Journal Of Climatology* **25**(15), 1965–1978.

Hoffmann, Irene (2011), « Livestock biodiversity and sustainability », *Livestock Science* **139**(1-2, SI), 69–79.

Holderegger, Rolf, Doris Herrmann, Bénédicte Poncet et al. (2008), « Land ahead : using genome scans to identify molecular markers of adaptive relevance », *Plant Ecology & Diversity* **1**(2), 273–283.

Holsinger, Kent E. et Bruce S. Weir (2009), « Genetics in geographically structured populations : defining, estimating and interpreting FST », *Nature Reviews Genetics* **10**(9), 639–650.

Hubby, JL et Richard C. Lewontin (1966), « A Molecular Approach To Study Of Genic Heterozygosity In Natural Populations .I. Number Of Alleles At Different Loci In Drosophila Pseudoobscura », *Genetics* **54**(2), 577–&.

Hudson, Richard R., Martin Kreitman et Montserrat Aguade (1987), « A Test Of Neutral Molecular Evolution Based On Nucleotide Data », *Genetics* **116**(1), 153–159.

Illumina Inc. (2012a), *BovineHD Genotyping BeadChip*.

– (2012b), *BovineSNP50 Genotyping BeadChip*.

ISO/IEC JTC1/SC22/WG21 (2011), *International Standard : The C++ Language*, rapp. tech. ISO/IEC 14882 :2011.

Jensen, Jeffrey D., Kevin R. Thornton, Carlos D. Bustamante et Charles F. Aquadro (2007), « On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations », *Genetics* **176**(4), 2371–2379.

Jones, Matthew, Brenna Forester, Ashley Teufel et al. (2013), « Integrating spatially explicit approaches to detect adaptive loci in a landscape genomics context », *Evolution*, Submitted.

Joost, Stéphane (2006), « The geographical dimension of genetic diversity : GIScience contribution for the conservation of animal genetic resources », thèse n^{o} 3454, Ecole Polytechnique Fédérale de Lausanne.

Joost, Stéphane, Aurélie Bonin, Michael W. Bruford et al. (2007), « A Spatial Analysis Method (SAM) to detect candidate loci for selection : towards a landscape genomics approach to adaptation », *Molecular Ecology* **16**(18), 3955–3969.

Joost, Stéphane, Aurélie Bonin, Pierre Taberlet et Régis Caloz (2008), « Un rôle pour la science de l'information géographique en écologie moléculaire », *Revue internationale de Géomatique* **18**(2), 215–237.

Joost, Stéphane et Michaël Kalbermatten (2010), *MatSAM Version 2Beta*, Laboratory of Geographic information systems, École polytechnique fédérale de Lausanne.

Joost, Stéphane, Michael Kalbermatten, Etienne Bezault et Ole Seehausen (2012), « Use of Qualitative Environmental and Phenotypic Variables in the Context of Allele Distribution Models : Detecting Signatures of Selection in the Genome of Lake Victoria Cichlids », in :

Data Production and Analysis in Population Genomics : Methods and Protocols, sous la dir. de François Pompanon et Aurélie Bonin, t. 888, Methods in Molecular Biology, Springer, 295–314.

Joost, Stéphane, Michaël Kalbermatten et Aurélie Bonin (2008), « Spatial Analysis Method (SAM) : a software tool combining molecular and environmental data to identify candidate loci for selection », *Molecular Ecology Ressources* **8**, 957–960.

Joost, Stéphane, Séverine Vuilleumier, Jeffrey D. Jensen et al. (2013), « Uncovering the genetic basis of adaptive change : on the intersection of landscape genomics and theoretical population genetics », *Molecular Ecology* **22**(14), 3659–3665.

Joshi, N.R., E.A. McLaughlin et Ralph W. Phillips (1957), *Les bovins d'Afrique : Types et races*, Rome : Food and Agriculture Organisation.

Kijas, James W., Johannes A. Lenstra, Ben Hayes et al. (2012), « Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection », *PLOS Biology* **10**(2).

Kimura, Motoo (1968), « Evolutionary Rate At Molecular Level », *Nature* **217**(5129), 624–&.

Landguth, Erin L. et Samuel A. Cushman (2010), « CDPOP : A spatially explicit cost distance population genetics program », *Molecular Ecology Ressources* **10**(1), 156–161.

Landguth, Erin L., Samuel A. Cushman et N. A. Johnson (2012), « Simulating natural selection in landscape genetics », *Molecular Ecology Ressources* **12**(2), 363–368.

Leempoel, Kevin, Jonathan Rolland, Ivo Widmer et al. (2013), « Comparing models of allele distribution to detect loci under selection from a genome scan », In preparation.

Lewontin, Richard C. (1973), « Population Genetics », *Annual Review of Genetics* **7**(1), PMID : 4593304, 1–17, eprint : http://www.annualreviews.org/doi/pdf/10.1146/annurev.ge.07.120173.000245.

Lewontin, Richard C. et J. L. Hubby (1966), « A Moleuclar Approach To Study Of Genic Heterozygosity In Natural Populations .II. Amount Of Variation And Degree Of Heterozygosity In Natural Populations Of Drosophila Pseudoobscura », *Genetics* **54**(2), 595–&.

Lewontin, Richard C. et Jesse Krakauer (1973), « Distribution Of Gene Frequency As A Test Of Theory Of Selective Neutrality Of Polymorphisms », *Genetics* **74**(1), 175–195.

Li, Jun Z., Devin M. Absher, Hua Tang et al. (2008), « Worldwide human relationships inferred from genome-wide patterns of variation », *Science* **319**(5866), 1100–1104.

Li, Junrui, Haipeng Li, Mattias Jakobsson et al. (2012), « Joint analysis of demography and selection in population genetics : where do we stand and where could we go ? », *Molecular Ecology* **21**(1), 28–44.

Lin, Xihong et Daowen Zhang (1999), « Inference in generalized additive mixed models by using smoothing splines », *Journal of the Royal Statistical Society Series B-Statistical Methodology* **61**(2), 381–400.

Lindgren, Finn, Havard Rue et Johan Lindstrom (2011), « An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic partial differential equation approach », *Journal of the Royal Statistical Society Series B-Statistical Methodology* **73**(Part 4), 423–498.

Liu, Lin, Yinhu Li, Siliang Li et al. (2012), « Comparison of Next-Generation Sequencing Systems », *Journal pf Biomedicine and Biotechnology*.

Luikart, Gordon, Phillip R. England, David Tallmon, Steve Jordan et Pierre Taberlet (2003), « The power of population genomics : from genotyping to genome typing », *Trends in Ecology and Evolution* **4**(12), 981–994.

Lundgren, Petra, Juan C. Vera, Lesa Peplow, Stephanie Manel et Madeleine J. H. van Oppen (2013), « Genotype - environment correlations in corals from the Great Barrier Reef », *BMC Genetics* **14**.

MacCallum, Iain, Dariusz Przybylski, Sante Gnerre et al. (2009), « ALLPATHS 2 : small genomes assembled accurately and with high continuity from short paired reads », *Genome Biology* **10**(10).

Malécot, Gustave (1948), *Les Mathématiques de l'Hérédité*, Paris : Masson.

Manel, Stéphanie, Cécile H. Albert et Nigel G. Yoccoz (2012), « Sampling in landscape genomics », in : *Data Production and Analysis in Population Genomics : Methods and Protocols*, sous la dir. de François Pompanon et Aurélie Bonin, t. 888, Methods in Molecular Biology, Springer, 3–12.

Manel, Stéphanie et Rolf Holderegger (2013), « Ten years of landscape genetics », *Trends in Ecology and Evolution*, pages.

Manel, Stéphanie, Stéphane Joost, Bryan K. Epperson et al. (2010), « Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field », *Molecular Ecology* **19**(17), 3760–3772.

Manel, Stéphanie, Bénédicte Poncet, Pierre Legendre, Felix Gugerli et Rolf Holderegger (2010), « Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina », *Molecular Ecology* **19**(17, SI), 3824–3835.

Manel, Stéphanie, Michael K. Schwartz, Gordon Luikart et Pierre Taberlet (2003), « Landscape genetics : combining landscape ecology and population genetics », *Trends in Ecology and Evolution* **18**(4), 189–197.

McDonald, John H. et Martin Kreitman (1991), « Adaptive protein evolution at the Adh locus in Drosophila », *Nature* **351**(6328), 652–654.

Meirmans, Patrick G. (2012), « The trouble with isolation by distance », *Molecular Ecology* **21**(12), 2839–2846.

Mendel, Gregor (1866), « Versuche über Pflanzenhybriden », *Verhandlungen des naturforschenden Vereines* **4**, 3–47.

Mingard, Patrick (2008), « Manipulation orientée objet de base de données en PHP », mém.de mast., École polytechnique fédérale de Lausanne.

Mirkena, T., G. Duguma, A. Haile et al. (2010), « Genetics of adaptation in domestic farm animals : A review », *Livestock Science* **132**(1-3), 1–12.

Mitton, J. B., Y. B. Linhart, J. L. Hamrick et J. S. Beckman (1977), « Observations On Genetic Structure And Mating System Of Ponderosa Pine In Colorado Front Range », *Theoretical and Applied Genetics* **51**(1), 5–13.

Mitton, Jeffry B., K. B. Sturgeon et M. L. Davis (1980), « Genetic Differentiation In Ponderosa Pine Along A Steep Elevational Transect », *Silvae Genetica* **29**(3-4), 100–103.

Mopper, S., Jeffry B. Mitton, T. G. Whitham, N. S. Cobb et K. M. Christensen (1991), « Genetic Differentiation And Heterozygosity In Pinyon Pine Associated With Resistance To Herbivory And Environmental-Stress », *Evolution* **45**(4), 989–999.

Moran, P. A. P. (1950), « Notes on Continuous Stochastic Phenomena », *Biometrika* **37**(1/2), pages.

Morgan, Thomas Hunt, A. H. Sturtevant, H. J. Muller et C.B. Bridges (1915), *The Mechanism of Mendelian Heredity*, New York : Henry Holt.

Morgenthaler, Stephan (2007), *Introduction à la statistique*, 3^{e} éd., Lausanne : Presses polytechniques et universitaires romandes.

– (2008), *Génétique statistique*, Paris : Springer.

Morin, Philip A., Gordon Luikart, Robert K. Wayne et SNP Workshop Grp (2004), « SNPs in ecology, evolution and conservation », *Trends in Ecology and Evolution* **19**(4), 208–216.

Mourant, Arthur E., Don Tills et Kazimiera Domaniewska-Sobczak (1976), « Sunshine and the geographical distribution of the alleles of the Gc system of plasma proteins », *Human Genetics* **33**(3), 307–314.

Neuenschwander, Samuel, Frédéric Hospital, Frédéric Guillaume et Jérôme Goudet (2008), « quantiNemo : an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation », *Bioinformatics* **24**(13), 1552–1553, eprint : http://bioinformatics.oxfordjournals.org/content/24/13/1552.full.pdf+html.

New, Mark, David Lister, Mike Hulme et Ian Makin (2002), « A high-resolution data set of surface climate over global land areas », *Climate Research* **21**, 1–25.

Nielsen, Rasmus, Carlos D. Bustamante, AG Clark et al. (2005), « A scan for positively selected genes in the genomes of humans and chimpanzees », *PLOS Biology* **3**(6), 976–985.

Nielsen, Rasmus, Ines Hellmann, Melissa Hubisz, Carlos Bustamante et Andrew G. Clark (2007), « Recent and ongoing selection in the human genome », *Nature Reviews Genetics* **8**(11), 857–868.

Ohta, Tomoko (1973), « Slightly Deleterious Mutant Substitutions In Evolution », *Nature* **246**(5428), 96–98.

Orr, H. Allen (2005), « The genetic theory of adaptation : a brief history », *Nature Reviews Genetics* **6**(2), 119–127.

– (2009), « Fitness and its role in evolutionary genetics », *Nature Reviews Genetics* **10**(8), 531–539.

Ortego, Joaquin, Maria P. Aguirre et Pedro J. Cordero (2012), « Landscape genetics of a specialized grasshopper inhabiting highly fragmented habitats : a role for spatial scale », *Diversity and Distributions* **18**(5), 481–492.

Nature editorial, (2013), « Overtaken by events. » *Nature* **497**(7451).

Patterson, Nick, Alkes L. Price et David Reich (2006), « Population structure and eigenanalysis », *PLOS Genetics* **2**(12), 2074–2093.

Paweletz, Neidhard (2001), « Walther Flemming : pioneer of mitosis research », *Nature Reviews Molecular Cell Biology* **2**(1), 72–75.

Pemstein, Daniel, Kevin M. Quinn et Andrew D. Martin (2011), « The Scythe Statistical Library : An Open Source C++ Library for Statistical Computation », *Journal of Statistical Software* **42**(12), 1–26.

Peter, C., Michael W. Bruford, T. Perez et al. (2007), « Genetic diversity and subdivision of 57 European and Middle-Eastern sheep breeds », *Animal Genetics* **38**(1), 37–44.

Poncet, Bénédicte, Doris Herrmann, Felix Gugerli et al. (2010), « Tracking genes of ecological relevance using a genome scan in two independent regional population samples of Arabis alpina », *Molecular Ecology* **19**(14), 2896–2907.

Pritchard, Jonathan K., Matthew Stephens et Peter Donnelly (2000), « Inference of population structure using multilocus genotype data », *Genetics* **155**(2), 945–959.

Purcell, Shaun (2009), *PLINK 1.07*, http://pngu.mgh.harvard.edu/purcell/plink/.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown et al. (2007), « PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses », *The American Journal of Human Genetics* **81**(3), 559–575.

R Core Team (2013), *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rebaudo, François, Arnaud Le Rouzic, Stéphane Dupas et al. (2013), « SimAdapt : an individual-based genetic model for simulating landscape management impacts on populations », *Methods in Ecology and Evolution* **4**(6), 595–600.

Rue, Havard, Sara Martino et Nicolas Chopin (2009), « Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations », *Journal of the Royal Statistical Society Series B-Statistical Methodology* **71**(Part 2), 319–392.

Sabeti, Pardis C., David E. Reich, John M. Higgins et al. (2002), « Detecting recent positive selection in the human genome from haplotype structure », *Nature* **419**(6909), 832–837.

Saiki, RK, DH Gelfand, S Stoffel et al. (1988), « Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase », *Science* **239**(4839), 487–491, eprint : http://www.sciencemag.org/content/239/4839/487.full.pdf.

Salakhutdinov, Ruslan et Andriy Mnih (2008), « Bayesian probabilistic matrix factorization using Markov chain Monte Carlo », in : *Proceedings of the 25th international conference on Machine learning*, ICML '08, Helsinki, Finland : ACM, 880–887.

Santos-del-Blanco, L., J. Climent, S. C. González-Martínez et J. R. Pannell (2012), « Genetic differentiation for size at first reproduction through male versus female functions in the widespread Mediterranean tree Pinus pinaster », *Annals of Botany* **110**(7), 1449–1460, eprint : http://aob.oxfordjournals.org/content/110/7/1449.full.pdf+html.

Schoville, Sean D., Aurélie Bonin, Olivier Francois et al. (2012), « Adaptive Genetic Variation on the Landscape : Methods and Cases », *Annual Review of Ecology, Evolution, and Systematics* **43**(1), 23–43, eprint : http://www.annualreviews.org/doi/pdf/10.1146/annurev-ecolsys-110411-160248.

Schrodt, Philip A. (2010), « Seven Deadly Sins of Contemporary Quantitative Political Analysis », in : *American Political Science Association 2010 Annual Meeting*, sous la dir. d'Andrea Campbell et Lisa Martin.

Schwartz, Michael K., Gordon Luikart, Kevin S. McKelvey et Samuel A. Cushman (2009), « Landscape Genomics : A Brief Perspective », in : *Spatial Complexity, Informatics and Wildlife Conservation*, sous la dir. de Springer, Cushman, Samuel A. et F. Huettmann, chap. 9, 165–174.

Schwartz, Michael K. et Kevin S. McKelvey (2009), « Why sampling scheme matters : the effect of sampling scheme on landscape genetic results », *Conservation Genetics* **10**(2), 441–452.

Slatkin, M (1991), « Inbreeding Coefficients And Coalescence Times », *Genetical Research* **58**(2), 167–175.

Sokal, Robert R. et Neal L. Oden (1978[a]), « Spatial Autocorrelation In Biology .1. Methodology », *Biological Journal of the Linnean Society* **10**(2), 199–228.

– (1978[b]), « Spatial autocorrelation in biology : 2. Some biological implications and four applications of evolutionary and ecological interest », *Biological Journal of the Linnean Society* **10**(2), 229–249.

Storey, J. D. et R. Tibshirani (2003), « Statistical significance for genomewide studies », *Proceedings of the National Academy of Sciences of the United States of America* **100**(16), 9440–9445.

Stroustrup, Bjarne (2013), *The C++ Programming Language*, 4^{e} éd., Pearson Education.

Stucki, Sylvie, Saif Agha, Meng-Hua Li et Stéphane Joost (2012), « Utilization of the Scythe C++ open source library for statistical geocomputation in livestock landscape genomics », in : *OGRS 2012 : Symposium proceedings of the Open Source Geospatial Research & Education Symposium*, sous la dir. d'Olivier Ertz, Stéphane Joost et Marj Tonini, Yverdon-les-Bains, Switzerland, 183–192.

Stutz, H. P. et Jeffry B. Mitton (1988), « Genetic-Variation In Engelmann Spruce Associated With Variation In Soil-Moisture », *Arctic and Alpine Research* **20**(4), 461–465.

Swynghedauw, Bernard (2008), *Biologie et génétiques moléculaires : Aide-mémoire*, 3^{e} éd., Paris : Dunod.

Taberlet, Pierre, Eric Coissac, Johan Pansu et François Pompanon (2011), « Conservation genetics of cattle, sheep, and goats », *Comptes Rendus Biologies* **334**(3), 247–254.

Taberlet, Pierre, Alessio Valentini, Hamid R. Rezaei et al. (2008), « Are cattle, sheep, and goats endangered species ? », *Molecular Ecology* **17**, 275–284.

Tajima, Fumio (1989), « Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. » *Genetics* **123**(3), 585–95, eprint : http://www.genetics.org/content/123/3/585.full.pdf+html.

The 1000 Genomes Project Consortium (2012), « An integrated map of genetic variation from 1,092 human genomes », *Nature* **491**(7422), 56–65.

The Bovine Genome Sequencing and Analysis Consortium, Christine G. Elsik, Ross L. Tellam et Kim C. Worley (2009), « The Genome Sequence of Taurine Cattle : A Window to Ruminant Biology and Evolution », *Science* **324**(5926), 522–528.

The UniProt Consortium (2013), « Update on activities at the Universal Protein Resource (UniProt) in 2013 », *Nucleic Acids Research* **41**(D1), D43–D47, eprint : http://nar.oxfordjournals.org/content/41/D1/D43.full.pdf+html.

Tipping, ME et CM Bishop (1999), « Probabilistic principal component analysis », *Journal of the Royal Statistical Society Series B-Statistical Methodology* **61**(3), 611–622.

Tobler, Waldo R. (1970), « A Computer Movie Simulating Urban Growth in the Detroit Region », *Economic Geography* **46**(2), 234–240.

Vitalis, R, K Dawson et P Boursot (2001), « Interpretation of variation across marker loci as evidence of selection », *Genetics* **158**(4), 1811–1823.

Vitalis, R., K. Dawson, P. Boursot et K. Belkhir (2003), « DetSel 1.0 : A Computer Program to Detect Markers Responding to Selection », *Journal of Heredity* **94**(5), 429–431, eprint : http://jhered.oxfordjournals.org/content/94/5/429.full.pdf+html.

Volis, Sergei (2008), « Population Genetics Research Progress », in : sous la dir. de Viktor T. Koven, Hauppauge, NY, USA : Nova Science Publishers, chap. Detection of signatures of positive selection in naturally occurring genetic variation.

Weinberg, Wilhelm (1908), « Über den Nachweis der Vererbung beim Menschen », *Hahreshefte des Vereins für vaterländische Naturkunde in Württemberg* **64**, 368–382.

Wright, Sewall (1943), « Isolation by distance », *Genetics* **28**(2), 114–138.

– (1931), « Evolution in Mendelian Populations », *Genetics* **16**(2), 97–159, eprint : http://www.genetics.org/content/16/2/97.full.pdf+html.

– (1932), « The roles of mutation, inbreeding, crossbreeding, and selection in evolution », in : *Proceedings of the Sixth International Congress of Genetics*, sous la dir. de Donald F. Jones, t. 1, 356–366.

– (1949), « The Genetical Structure Of Populations », *Annals of Eugenics* **15**(1), 323–354.

Glossaire

Les références indiquent la source de la définition ou des ouvrages ou articles fournissant des explications plus détaillées.

allèle (Morgenthaler, 2008)

- Copie d'un gène ou d'un marqueur génétique. (Chaque organisme diploïde possède deux allèles de chaque gène, sauf pour les chromosomes sexuels.)
- Variante d'un gène ou d'un marqueur génétique. (Un individu hétérozygote à un locus possède deux allèles différents pour ce locus.)

confounding factor voir facteur de confusion (ou facteur parasite).

dérive génétique Variation des fréquences alléliques d'une génération à l'autre au sein d'une population de taille finie sous l'effet de la transmission aléatoire des allèles. Cet effet augmente lorsque la taille d'une population diminue. (*genetic drift*, Orr, 2009 ; Luikart et al., 2003)

facteur de confusion (ou facteur parasite) Facteur non-mesuré (ou immesurable) dans une expérience qui déforme l'association de deux autres variables et peut mener à de fausses conclusions quand à l'influence d'un facteur sur un phénomène. Dans le contexte de la génétique des populations, l'effet de l'histoire démographique d'une population peut être facilement confondu avec une signature de sélection et peut mener à de fausses détections. (*confounding factor*, Nielsen et al., 2007 ; Li et al., 2012)

fitness voir valeur sélective.

flux de gènes Transfert d'allèles ou de gènes entre deux entités, par exemple lors de la migration d'un individu d'une population vers une autre. (*gene flow*, Manel et Holderegger, 2013)

gene flow voir flux de gènes.

genetic drift voir dérive génétique.

génome Ensemble du matériel génétique d'un individu ou d'une cellule, dont il constitue le génotype (Swynghedauw, 2008).

génomique des populations Étude simultanée de nombreux loci ou régions du génome afin d'expliquer les rôles des processus évolutifs (comme les mutations, la dérive génétique, les flux de gènes et la sélection naturelle) qui influencent les variations génétiques d'un individu à l'autre ou d'une population à l'autre. (*Population Genomics*, Luikart et al., 2003)

génomique environnementale Domaine de la génétique environnementale spécialisé dans la détection de régions du génome potentiellement soumises à la sélection naturelle. Cette discipline se base sur des modèles corrélatifs associant directement des données génétiques et des variables environnementales pour estimer l'influence de l'habitat sur le patrimoine génétique d'un organisme. (*Landscape Genomics*, Joost et al., 2007 ; Joost et al., 2013)

génotypage Acte technique permettant de déterminer un génotype donné, soit pour un locus, soit pour un ensemble de loci (Swynghedauw, 2008).

génotype (Henry et Gouyon, 2008)

- Constitution génétique d'un individu.
- Composition allélique du locus (ou des loci) étudié(s) chez un individu.

génétique des populations Discipline étudiant la variation dans le temps (et parfois dans l'espace) du patrimoine génétique de populations d'organismes sous l'effet des forces évolutives (comme les mutations, la dérive génétique, les flux de gènes et la sélection naturelle). (*Population Genetics*, Henry et Gouyon, 2008)

génétique environnementale Discipline étudiant l'interaction entre les particularités de l'environnement où vit un organisme et les processus micro-évolutifs à l'oeuvre (comme les flux de gènes, la dérive génétique ou la sélection). La connaissance de la position géographique des individus étudiés permet de caractériser leurs habitats, ce qui offre la possibilité de comparer l'environnement d'un organisme avec son patrimoine génétique. (*Landscape Genetics*, Manel et al., 2003 ; Schoville et al., 2012)

Landscape Genetics voir génétique environnementale.

Landscape Genomics voir génomique environnementale.

locus Position spécifique sur le génome (Henry et Gouyon, 2008).

migration En génétique, déplacement d'un individu d'une population vers une autre (Henry et Gouyon, 2008).

mutation Modification spontanée et héréditaire du génome, par ex. une substitution d'un nucléotide par un autre à un locus spécifique ou le déplacement (translocation) d'un fragment d'ADN dans le génome. Une mutation peut mener à l'apparition d'un nouvel allèle. (Morgenthaler, 2008 ; Henry et Gouyon, 2008)

- **mutation bénéfique (ou adaptative)** qui favorise son porteur face à la sélection naturelle (augmentation de la valeur sélective, *beneficial mutation*).
- **mutation délétère (ou dommageable)** qui handicape son porteur face à la sélection naturelle (baisse de la valeur sélective, *deleterious mutation*).

mutation neutre sans influence sur la valeur sélective de l'individu qui la porte (*neutral mutation*).

phénotype Expression du génotype dans un milieu donné, caractéristique observable d'un individu (Henry et Gouyon, 2008).

polymorphisme nucléotidique Emplacement de l'ADN où le nucléotide présent (A, C, T ou G) peut varier d'un individu à l'autre ou entre les deux chromosomes homologues d'un individu. (*Single Nucleotide Polymorphism*, SNP, Morin et al., 2004)

Population Genetics voir génétique des populations.

Population Genomics voir génomique des populations.

Single Nucleotide Polymorphism voir polymorphisme nucléotidique.

SNP *Single Nucleotide Polymorphism*, voir polymorphisme nucléotidique.

sélection Processus entraînant une reproduction supérieure ou inférieure chez certains individus du fait de leur génotype. La sélection peut être naturelle ou exercée par l'homme pour l'élevage ou l'agriculture (Henry et Gouyon, 2008).

séquençage Déchiffrage d'un fragment d'ADN nucléotide par nucléotide, ce fragment peut être un gène, un chromosome ou un génome entier (Swynghedauw, 2008).

valeur sélective Quantité proportionnelle au nombre moyen de descendants viables et fertiles d'un individu. La valeur sélective est une mesure de l'avantage ou du handicap fourni par un génotype face à la sélection naturelle. (*fitness*, Orr, 2005 ; Orr, 2009)

Curriculum Vitae

Personal information

Name	Sylvie Stucki
Work adress	EPFL ENAC IIE LASIG, Station 18, CH-1015 Lausanne
Home adress	Ch. de Champ-Fleuri 12, CH-1022 Chavannes-près-Renens
Birth	20^{th} July 1984 in Aigle (Switzerland)
Nationality	Swiss
Status	Single
E-mail	sylvie.stucki@epfl.ch

Education and degrees

2007-2009	Master of Science MSc in Physics at EPFL. *Computation of liquid water polarisability using the Wannier functions method.* (Supervisor: Prof Alfredo Pasquarello)
2003-2007	Bachelor of Science, section of Physics at EPFL.
2000-2003	Federal matura at Gymnasium of Burier option Physics and Application of mathematics.

Work, research and teaching experience

2009-present	PhD at École polytechnique fédérale de Lausanne, Laboratory of Geographic Information Systems (LaSIG) *Spatial analysis of whole genome diversity and HPC geocomputational tools to identify adaptive genomic regions* (Supervisors: Prof François Golay and Dr. Stéphane Joost)
2010-present	NextGen project *Next generation methods to preserve farm animal biodiversity by optimizing present and future breeding options* (Coordinator: Pierre Taberlet) Landscape genomics study of local adaptation and disease resistance in livestock, bioinformatics developments
2009-present	Teaching assistant in LaSIG (Spatial analysis, Geovisualisation and Geocomputation)
2006-2008	Teaching assistant in computer science, language C++

Conference papers

Stucki Sylvie, Pablo Orozco-terWengel, Licia Colli, Fredrick Kabi, Charles Masembe, Vincent Muwanika, Riccardo Negrini, Michael W. Bruford, Stéphane Joost and the NEXTGEN Consortium, 2013, *Samβada in Uganda: landscape genomics study of traditional cattle breeds with a large SNP dataset*, in "Proceedings of the International Association for Landscape Ecology Conference", (in press).

Stucki Sylvie, Saif Agha, Meng-Hua Li and Stéphane Joost, 2012, *Utilization of the Scythe C++ open source library for statistical geocomputation in livestock landscape genomics*, in "Proceedings of the Open Source Geospatial Research & Education Symposium (OGRS 2012), Olivier Ertz, Stéphane Joost and Marj Tonini (eds.), Lausanne.

Computer skills

Programming	C++, R, SQL and Fortran
Software	Manifold & QuantumGIS (Geographic Information System)
	OpenGeoda (Spatial analysis)

Languages

French	mother tongue
English	good knowledge
German	basic knowledge